Probability Theory

Volker Betz

1. Basic definitions and facts

In this section, we recall the definitions and statements from the lecture 'Einführung in die Stochastik' that will be relevant for this course. Another purpose of this repetition is to make you familiar with the English terminology.

(1.1) Definition

Let Ω be a non-empty set. A collection of subsets $\mathcal{F} \subset \mathcal{P}(\Omega)$ is a *σ-algebra (σ-field)* over Ω if

a) $\emptyset \in \mathcal{F}$,

b) $A \in \mathcal{F} \Rightarrow A^c := \Omega \backslash A \in \mathcal{F}$,

c) $A_i \in \mathcal{F} \, \forall i \in \mathbb{N} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$.

$(\Omega, \mathcal{F})$ is called a *measurable space.*
Remark: You should check that the conditions imply $\bigcap_{i \in \mathbb{N}} A_i \in \mathcal{F}$ in the situation of c).

(1.2) Definition

Let $\mathcal{F}$ be a σ-algebra over Ω. A *measure* is a map $\mu : \mathcal{F} \to \mathbb{R}_0^+$ with

a) $\mu(\emptyset) = 0$.

b) $\mu\big(\bigcup_{i \in \mathbb{N}} A_i\big) = \sum_{i \in \mathbb{N}} \mu(A_i)$ for all **disjoint** collections of $A_i \in \mathcal{F}$.

If $\mu(\Omega) = 1$, then μ is called a *probability measure.* In this case we often use the letter $\mathbb{P}$ instead of μ and call $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space.*

(1.3) Lemma

Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$, A, B, $(A_i) \in \mathcal{F}$.

a) $A \subset B \Rightarrow \mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B \setminus A) \geqslant 0$; ('Monotonicity');

b) $A \subset \bigcup_{i \in \mathbb{N}} A_i \implies \mathbb{P}(A) \leqslant \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$; ('Subadditivity');

c) $A_i \nearrow A \implies \mathbb{P}(A_i) \nearrow \mathbb{P}(A)$; ('Continuity from below');

d) $A_i \searrow A \implies \mathbb{P}(A_i) \searrow \mathbb{P}(A)$; ('Continuity from above').

Notational Remarks:
a) Above, $A_i \nearrow A$ means that $A_i \subset A_{i+1}$ for all i and $\bigcup_i A_i = A$. $A_i \searrow A$ means $A_{i+1} \subset A_i$ for all i and $\bigcap_i A_i = A$.
b) We will often write $\nearrow$ instead of $\nearrow_{i \to \infty}$ if it is clear what limit is taken. Similarly, we will often write $\bigcup_i$ instead of $\bigcup_{i \in \mathbb{N}}$ if it is clear what set the i are from.

(1.4) Most important elementary examples

a) Ω countable set, $\mathcal{F} = \mathbb{P}(\Omega)$, $p : \Omega \to \mathbb{R}_0^+$ with $\sum_{\omega \in \Omega} p(\omega) = 1$. Then

$$\mathbb{P}(A) := \sum_{\omega \in A} p(\omega) = \sum_{\omega \in \Omega} 1_A(\omega) p(\omega)$$

defines a probability measure. $1_A(\omega) = 1$ if $\omega \in A$ and $= 0$ otherwise, is called the *indicator function* of A. p is called the *probability weight function.* All probability measures on countable spaces are of the above form.
b) $\Omega = \mathbb{R}^d$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^d) \equiv \mathcal{B}^d$ the Borel-σ-algebra, $f : \mathbb{R}^d \to \mathbb{R}_0^+$ a measurable function with $\int_{\mathbb{R}^d} f(x)\mathrm{d}x = 1$. Then

$$\mathbb{P}(A) := \int_A f(x)\mathrm{d}x = \int_{\mathbb{R}^d} f(x) 1_A(x)\,\mathrm{d}x \quad (A \in \mathcal{B}^d)$$

defines a probability measure on $\mathbb{R}^d$. f is called *Lebesgue-density* of $\mathbb{P}$. Not all probability measures on $\mathbb{R}^d$ can be written in the form above, but many important ones can.

(1.5) Product spaces, product measures, generated σ-algebras

Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. We define the cartesian product

$$\Omega_1 \times \Omega_2 \times \ldots \times \Omega_n \equiv \prod_{i=1}^n \Omega_i = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i \forall i\},$$

and

$$\mathcal{F} := \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \ldots \otimes \mathcal{F}_n \equiv \bigotimes_{i=1}^n \mathcal{F}_i$$

to be the smallest σ-algebra over $\prod_{i=1}^n \Omega_i$ that contains all *measurable rectangles*

$$A_1 \times \ldots \times A_n := \{(\omega_1, \ldots, \omega_n) : \omega_i \in A_i \forall i\}$$

with $A_i \in \mathcal{F}_i$ for all i. If we construct a σ-algebra as the smallest σ-algebra containing a certain collection $\mathcal{G} \subset \mathcal{P}(\Omega)$ of subsets, we say that it is the σ-algebra *generated by* $\mathcal{G}$ and write $\sigma(\mathcal{G})$.
If $\mathbb{P}_i$ is a probability measure on $(\Omega_i, \mathcal{F}_i)$ for all i, then there is a unique probability measure $\mathbb{P} := \mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \ldots \otimes \mathbb{P}_n$ on Ω so that for all measurable rectangles $A = A_1 \times \ldots \times A_n$ the equality

$$\mathbb{P}(A) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \cdots \mathbb{P}_n(A_n)$$

holds. This measure is called the *product* of the $\mathbb{P}_i$. As a main example, think of the n-dimensional Lebesgue measure.

(1.6) Random variables, distributions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(\Omega', \mathcal{F}')$ be a measurable space, i.e. a set Ω' with a σ-algebra $\mathcal{F}'$ over Ω'. A function $X : \Omega \to \Omega'$ is called *random variable (RV)* if it is measurable, i.e. if

$$X^{-1}(A') \equiv \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{F}$$

holds for all $A' \in \mathcal{F}'$. The *distribution* of a RV X is the measure $\mathbb{P}_X$ on Ω' so that

$$\mathbb{P}_X(A') = \mathbb{P}(X^{-1}(A')) \quad \text{for all } A' \in \mathcal{F}'.$$

We also say that $\mathbb{P}_X$ is the *image of* $\mathbb{P}$ *under* X, or the *pushforward of* $\mathbb{P}$ *under* X. When $\mathbb{P}_X = \mu$ for some probability measure μ, we also sometimes write $X \sim \mu$ and say that X is distributed like μ. Example: $X \sim \mathcal{N}(0,1)$. Finally, we sometimes write

$$\mathbb{P}_X = \mathbb{P} \circ X^{-1}$$

(1.7) Example

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

a) $A \in \mathcal{F}$, $X = 1_A$, then X is a real-valued RV and for $B \in \mathcal{B}$,

$$\mathbb{P}_X(B) = \begin{cases} \mathbb{P}(A) & \text{if } 1 \in B, 0 \notin B, \\ 1 - \mathbb{P}(A) & \text{if } 0 \in B, 1 \notin B, \\ 1 & \text{if } 0 \in B, 1 \in B, \\ 0 & \text{otherwise.} \end{cases}$$

More compactly, for $B \in \mathcal{B}$, $\mathbb{P}_X(B) = \mathbb{P}(A)\delta_1(B) + (1 - \mathbb{P}(A))\delta_0(B)$, where δ_x is the *Dirac measure* in the point $x \in \mathbb{R}$.

b) $A_i \in \mathcal{F}$, $A_i \cap A_j = \emptyset$ for all $i \neq j$, $X = \sum_{i=1}^n \alpha_i 1_{A_i}$. Then X is a real-valued RV, and you should convince yourself that

$$\mathbb{P}_X(\cdot) = \sum_{i=1}^n \mathbb{P}(A_i)\delta_{\alpha_i}(\cdot) + (1 - \mathbb{P}(\bigcup_{i=1}^n A_i))\delta_0(\cdot).$$

The last term can be dropped if we assume that $\bigcup_{i=1}^n A_i = \Omega$. In this case, X is called *elementary RV*.

(1.8) Distribution functions

Let X be a real-valued RV. The function

$$F : \mathbb{R} \to [0,1], \quad x \mapsto \mathbb{P}(X \leqslant x) = \mathbb{P}_X((-\infty, x])$$

is called *(cumulative) distribution function* of X. It uniquely determines the distribution of X (but not X itself).

(1.9) Equality in Distribution

Let X, Y be RVs. We say that X and Y have the same distribution, and write $X \overset{d}{=} Y$, if $\mathbb{P}_X = \mathbb{P}_Y$. Note that $X \overset{d}{=} Y$ is a much weaker statement than $X = Y$.

(1.10) Examples for important non-real-valued RVs

A RV $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega', \mathcal{F}')$ is also called a
random vector if $\Omega' = \mathbb{R}^d$, $\mathcal{F}' = \mathcal{B}^d$, $d > 1$;
random permutation if $\Omega' = \mathcal{S}(N) = \{\pi : \{1, \dots, N\} \to \{1, \dots, N\}, \pi \text{ bijective}\}$, $\mathcal{F}' = \mathcal{P}(\mathcal{S}_N)$.
random continuous function, or random path if $\Omega' = C(\mathbb{R}_0^+, \mathbb{R}^d)$. The σ-algebra is the one generated by the point evaluations (not relevant in this lecture).

(1.11) Testing for measurability

Measurability of a random variable needs only be tested on a collection of sets generating the σ-algebra on the target space of X. In symbols: Let $\mathcal{F}'$ be a σ-algebra, and let $\mathcal{G}$ be a collection of sets with $\sigma(\mathcal{G}) = \mathcal{F}'$. A function X is measurable from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$ if and only if $X^{-1}(A') \in \mathcal{F}$ for all $A' \in \mathcal{G}$.

(1.12) Generated σ-algebra

Let X be a RV from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$. The σ-algebra (!) $\sigma(X) := \sigma(\{X^{-1}(A) : A \in \mathcal{F}'\})$ is the smallest σ-algebra over Ω with respect to which X is measurable. It is called *the σ-algebra generated by* X and can be much smaller than $\mathcal{F}$. Try to find the σ-algebra generated by the RVs in Example (1.7).

(1.13) New RVs from old

a) If $X : \Omega \to \Omega'$, $Y : \Omega' \to \Omega''$ are RVs, then also $Y \circ X : \Omega \to \Omega''$ is a RV.

b) If X_i, $i = 1, \dots, n$, are real-valued RVs and $f : \mathbb{R}^n \to \mathbb{R}^m$ is measurable, then $f(X_1, \dots, X_n) : \Omega \to \mathbb{R}^m$ is a RV. This works the same with more general target spaces.

c) Special case of b): $\sum_{i=1}^n X_i$ is a RV.

d) Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of RVs. Then $\omega \mapsto \inf_i X_i(\omega)$ and $\omega \mapsto \sup_i X_i(\omega)$ are RVs. The same for $\omega \mapsto \limsup_{i \to \infty} X_i(\omega)$ and $\omega \mapsto \liminf_{i \to \infty} X_i(\omega)$.

e) In the situation of d), $\Omega_0 := \{\omega \in \Omega : \lim_{n \to \infty} X_i(\omega) \text{ exists}\}$ is measurable, and

$$1_{\Omega_0}(\omega) \lim_{n \to \infty} X_n(\omega) := 1_{\Omega_0}(\omega) \limsup_{n \to \infty} X_n(\omega) = 1_{\Omega_0}(\omega) \liminf_{n \to \infty} X_n(\omega)$$

is a RV.

Exercise: check that all the statements so far also hold for the probability space $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ with the σ-algebra generated by the intervals $[-\infty, a)$ and $(b, \infty]$ for $a, b \in \mathbb{R}$.

(1.14) Expected Value

a) Discrete probability spaces: $\Omega = \{\omega_i : i \in \mathbb{N}\}$;
A RV $X : \Omega \to \mathbb{R}^d$ is *integrable* if $\sum_{i\in\mathbb{N}} \|X(\omega_i)\| \mathbb{P}(\{\omega_i\}) < \infty$, where $\|x\|$ is any (e.g. the euclidean) norm on $\mathbb{R}^d$. In this case,

$$\mathbb{E}(X) := \sum_{i\in\mathbb{N}} X(\omega_i)\mathbb{P}(\{\omega_i\})$$

is the *expected value* of X with respect to $\mathbb{P}$ (defined component-wise).

b) Lebesgue densities: $\Omega = \mathbb{R}^n$, $\mathbb{P}(A) = \int_A \rho(y)\mathrm{d}y$ for $A \in \mathcal{B}^n$, with some density $\rho \geqslant 0$, $\int \rho(y)\,\mathrm{d}y = 1$.
A RV $X : \mathbb{R}^n \to \mathbb{R}^d$ is *integrable* if $\int \rho(y)\|X(y)\|\,\mathrm{d}y < \infty$, and in this case,

$$\mathbb{E}(X) := \int \rho(y)X(y)\,\mathrm{d}y$$

is the expected value of X with respect to the measure $\mathbb{P} = \rho(y)\,\mathrm{d}y$ (defined component-wise) .

c) General case, positive RV: $(\Omega, \mathcal{F}, \mathbb{P})$ general, $X : \Omega \to \mathbb{R}_0^+$ a RV. Then for each $n \in \mathbb{N}$, let

$$X_n(\omega) = \min\Big\{\frac{1}{n}\lfloor nX(\omega)\rfloor, n\Big\}.$$

Then, $\Omega_n := \{\omega_j : 0 \leqslant j \leqslant n^2\}$ with 'points' $\omega_j := X_n^{-1}(j/n) \subset \Omega$, for $0 \leqslant j \leqslant n^2$, and probability measures $\mathbb{P}_n(\{\omega_j\}) = \mathbb{P}\big(X_n^{-1}(\{j/n\})\big)$ is a discrete probability space (situation a), and we define

$$\mathbb{E}(X_n) := \mathbb{E}_{\mathbb{P}_n}(X_n) := \sum_{k=0}^{n^2} \frac{k}{n}\mathbb{P}_n(\{\omega_k\}) = \sum_{k=0}^{n^2} \frac{k}{n}\mathbb{P}(X_n^{-1}(\{k/n\})).$$

X is called *integrable* if $\mathbb{E}(X) := \lim_{n\to\infty} \mathbb{E}(X_n)$ exists, and $\mathbb{E}(X)$ is then called *expected value* of X. If $\mathbb{E}(X) = \infty$, then X is not called integrable, but we still say that the expected value exists and is $+\infty$.

d) General case, general RV: For real valued RVs, X is called *integrable* if positive part X_+ and negative part X_- are integrable in the sense of c). Then $\mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-)$.
For $\mathbb{R}^d$-valued RVs, X is integrable if the positive RV $\omega \mapsto \|X(\omega)\|$ is integrable. Equivalently, if each component is integrable. $\mathbb{E}(X)$ is then defined component-wise.

e) Properties of the expected value:

Linearity: If X, Y are integrable, then also $X + \alpha Y$ $(\alpha \in \mathbb{R})$ is integrable, and $\mathbb{E}(X + \alpha Y) = \mathbb{E}(X) + \alpha\mathbb{E}(Y)$.

Monotonicity: If $X(\omega) \geqslant Y(\omega)$ for all ω, then $\mathbb{E}(X) \geqslant \mathbb{E}(Y)$ if the expected values exist.
If $X(\omega) \geqslant Y(\omega)$ for all ω **and** $\mathbb{E}(X) = \mathbb{E}(Y)$, then $\mathbb{P}(X \neq Y) = 0$. In this case, we say that $X = Y$ *almost surely.*

(1.15) Variance

For a real-valued RV X with $\mathbb{E}(X) < \infty$, we define the *variance* of X as

$$\mathbb{V}(X) \equiv \mathrm{Var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

The variance can be infinite. We have that $\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$.

(1.16) Computing expected values

a) Let X be a $\mathbb{R}^d$-valued RV on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $f : \mathbb{R}^d \to \mathbb{R}^n$ be a measurable function so that $f(X)$ is integrable. Then

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}^d} f(y) \mathbb{P}_X(\mathrm{d}y),$$

where $\mathbb{P}_X$ is the image measure of $\mathbb{P}$ under X.
b) Special case: $d = n, f(x) = x$, then

$$\mathbb{E}(X) = \int_{\mathbb{R}^d} y \, \mathbb{P}_X(\mathrm{d}y).$$

If $d = n = 1$ and $f(x) = x^m$, then the *m-th moment* of X is defined as $\mathbb{E}(X^m)$ and can be calculated via

$$\mathbb{E}(X^m) = \int y^m \, \mathbb{P}_X(\mathrm{d}y), \quad m \geqslant 1.$$

c) Special case: Gaussian measure. If X is a Gaussian RV with expected value μ and variance σ^2, then $\mathbb{P}_X$ has the Lebesgue-density

$$\rho(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-\frac{1}{2\sigma^2}|y-\mu|^2} \, , \; y \in \mathbb{R}.$$

Then,

$$\mathbb{E}((X - \mu)^n) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (y - \mu)^n \, \mathrm{e}^{-\frac{1}{2\sigma^2}|y-\mu|^2} \, \mathrm{d}y = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ (n-1)!!\sigma^n & \text{if } n \text{ is even.} \end{cases}$$

Here $(n-1)!! = (n-1) \cdot (n-3) \cdots 3 \cdot 1$. Also, $\mathbb{E}(X^n)$ can be calculated by expanding $X^n = (X - \mu + \mu)^n = \sum_{k=0}^n \binom{n}{k} \mu^k (X - \mu)^{n-k}$ and using linearity.

d) A very useful formula for calculating expected values is the following: if X is a real-valued RV, $X \geqslant 0$, and if $f : \mathbb{R}^+ \to \mathbb{R}^+$ is differentiable and monotone increasing, then

$$\mathbb{E}(f(X)) = f(0) + \int_0^\infty f'(y) \, \mathbb{P}(X > y) \mathrm{d}y.$$

(Proof: the integral on the right hand side equals

$$\int_0^\infty \mathrm{d}y \, f'(y) \int_\Omega \mathbb{P}(\mathrm{d}\omega) 1_{\{X(\omega) > y\}} = \int \mathbb{P}(\mathrm{d}\omega) \int_0^\infty \mathrm{d}y \, f'(y) 1_{\{X(\omega) > y\}},$$

by Fubini's theorem which holds since $f' \geqslant 0$. The inner integral in the last expression on the right hand side above equals $\int_0^{X(\omega)} f'(y)\mathrm{d}y = f(X(\omega)) - f(0)$. This shows the claim.)

e) A frequently useful special case of the above equation is

$$\mathbb{E}(X^p) = \int p\, y^{p-1}\, \mathbb{P}(X > y)\, \mathrm{d}y.$$

(1.17) Inequalities for expected values

a) **Jensen:** If X is a real-valued RV, $\phi : \mathbb{R} \to \mathbb{R}$ is convex and both X and $\phi(X)$ are integrable, then

$$\mathbb{E}(\phi(X)) \geqslant \phi(\mathbb{E}(X)).$$

b) **Hölder:** For $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathbb{E}(|XY|) \leqslant \|X\|_p \|Y\|_q \qquad \text{where } \|X\|_p := \mathbb{E}(|X|^p)^{1/p}, \|Y\|_q := \mathbb{E}(|X|^q)^{1/q}.$$

c) **Cauchy-Schwarz:** Hölder with $p = q = 2$, i.e.

$$\mathbb{E}(|XY|)^2 \leqslant \mathbb{E}(|X|^2)\, \mathbb{E}(|Y|^2).$$

c) **Chebyshev-Markov:** Let X be a real-valued RV, $f : \mathbb{R} \to \mathbb{R}^+$ measurable and $A \in \mathcal{B}$. Then

$$\inf\{f(x) : x \in A\} \mathbb{P}(X \in A) \leqslant \mathbb{E}\left(f(X) 1_{\{X \in A\}}\right) \leqslant \mathbb{E}(f(X)).$$

(Proof: $\inf\{f(x) : x \in A\} 1_{\{X(\omega) \in A\}} \leqslant f(X(\omega)) 1_{\{X(\omega)\in A\}}$ for all ω. Now take expectation.)
Special cases:

$f(x) = x^2, a > 0, A = \mathbb{R} \backslash (-a, a) : \quad a^2 \mathbb{P}(|X| \geqslant a) \leqslant \mathbb{E}(X^2),$

$f(x) = x^2, X = |Z - \mathbb{E}(Z)|$ for some RV $Z, a > 0, A = [a, \infty) : \quad a^2 \mathbb{P}(|Z - \mathbb{E}(Z)| \geqslant a) \leqslant \mathbb{V}(Z).$

(1.18) Exchange of limit and integration

Let (X_n) be a sequence of RVs.

a) **Fatou:** $X_n \geqslant 0 \implies \liminf_{n\to\infty} \mathbb{E}(X_n) \geqslant \mathbb{E}(\liminf_{n\to\infty} X_n)$.

b) **Monotone Convergence:** If $0 \leqslant X_n \nearrow X$, then $\mathbb{E}(X_n) \nearrow \mathbb{E}(X)$.

c) **Dominated Convergence:** If $X_n \to X$ almost surely (a.s.), and if $|X_n| \leqslant Y$ a.s. for some integrable RV Y, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

(1.19) Independence

a) Two events (= measurable sets) A, B are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

We write $A \perp\!\!\!\perp B$.

b) Two collections $\mathcal{F}$ and $\mathcal{G}$ of sets are *independent* if

$$\forall A \in \mathcal{F}, \forall B \in \mathcal{G} : \quad \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

We write $\mathcal{F} \perp\!\!\!\perp \mathcal{G}$. The most important special case is when $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras.
c) Two RVs $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega', \mathcal{F}')$ are *independent* if

$$\forall A, B \in \mathcal{F}' : \quad \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B).$$

We write $X \perp\!\!\!\perp Y$.
d) Fact:

$X \perp\!\!\!\perp Y \Leftrightarrow \sigma(X) \perp\!\!\!\perp \sigma(Y) \Leftrightarrow \mathbb{E}\big(f(X)g(Y)\big) = \mathbb{E}\big(f(X)\big)\, \mathbb{E}\big(g(Y)\big) \quad \forall f, g$ bounded and measurable.

e) A family $(A_i)_{i \in I}$ of sets $A_i \subset \mathcal{F}$, with some index set I, is *independent* (or: an independent family of sets) if for all finite subsets $\{i_1, \dots, i_n\} \subset I$, we have that

$$\mathbb{P}\left(\bigcap_{k=1}^{n} A_{i_k} \right) = \prod_{k=1}^{n} \mathbb{P}(A_{i_k}).$$

f) A family of σ-algebras $\mathcal{G}_i$, $i \in I$, is an *independent family of σ-algebras* (or: independent) if e) holds for each choice of sets A_i with $A_i \in \mathcal{G}_i$.

g) A family of RVs X_i, $i \in I$, is called *independent family of RVs* (or: independent) if their σ-algebras $\mathcal{G}_i = \sigma(X_i)$ are independent.

(1.20) The π-λ-Theorem

a) A π-system is a collection $\mathcal{A} \subset \mathcal{P}(\Omega)$ of sets with the property that

$$A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}.$$

b) A λ-system is a collection $\mathcal{L} \subset \mathcal{P}(\Omega)$ of sets with the property that
(i): $\Omega \in \mathcal{L}$.
(ii): $A, B \in \mathcal{L}, B \subset A \implies A \setminus B \in \mathcal{L}$.
(iii): $A_n \in \mathcal{L}, n \in \mathbb{N}$, with $A_n \nearrow A \implies A \in \mathcal{L}$.

c) **Theorem:** Let $\mathcal{A}$ be a π-system and $\mathcal{L}$ be a λ-system. If $\mathcal{A} \subset \mathcal{L}$, then also $\sigma(\mathcal{A}) \subset \mathcal{L}$. An important special case is when $\mathcal{L}$ is itself a σ-algebra.

d) **Consequence:** If $\mathcal{A}_1, \dots, \mathcal{A}_n$ are π-systems and $\mathcal{A}_1, \dots, \mathcal{A}_n$ are independent, then $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent. (proof: exercise!)

e) **Application:** For a real-valued RV X, the system $\{X^{-1}((-\infty, c)) : c \in \mathbb{R}\}$ is a π-system. (Reason: the inverse image map preserves all set operations!) Therefore, two real-valued RVs X and Y are independent if and only if $\mathbb{P}(X < a, Y < b) = \mathbb{P}(X < a)\mathbb{P}(Y < b)$ for all $a, b \in \mathbb{R}$.

(1.21) Independence and Distribution

a) **Theorem:** If $X_1, \dots, X_n$ are RVs and $\mathbb{P}_{X_i} = \mu_i$, then the X_i are independent if and only if

$$\mathbb{P}_{(X_1, \dots X_n)} = \mu_1 \otimes \dots \otimes \mu_n \quad \text{(product measure!)}.$$

b) Consequence: Let X_i, $i = 1, \dots, n$, be independent real-valued RVs and μ_i probability measures on $\mathbb{R}$ with $X_i \sim \mu_i$ for all i. If $h : \mathbb{R}^n \to \mathbb{R}$ is either positive or has $\mathbb{E}(h(X_1, \dots, X_n)) <$

∞, then

$$\mathbb{E}(h(X_1,\dots,X_n)) = \int_{\mathbb{R}^n} h(x_1,\dots,x_n)\mu_1(\mathrm{d}x_1)\cdots\mu_n(\mathrm{d}x_n).$$

c) Special case of b): $h(x_1,\dots,x_n) = \prod_{i=1}^n f_i(x_i)$, then

$$\mathbb{E}\left(\prod_{i=1}^n f_i(X_i)\right) = \prod_{i=1}^n \mathbb{E}(f_i(X_i))$$

d) For $f_i(x_i) = x_i$ and $n = 2$ in c), we get

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \quad (*).$$

RVs having the property $(*)$ are called *uncorrelated*. As we have seen, independence implies uncorrelation. The converse is usually not true.

e) Notable exception to the last statement: if X and Y both have a Gaussian distribution, then X, Y uncorrelated implies X, Y independent.

(1.22) Sums of Random variables, convolution

a) If X and Y are real RVs, and if (X,Y) has a Lebesgue-density $(x,y) \mapsto \rho(x,y)$, then $X+Y$ has a Lebesgue-density $\int_{-\infty}^{\infty} \rho(x-y,y)\mathrm{d}y$. We write

$$Z = X + Y \sim \Big(\int_{-\infty}^{\infty} \rho(z-y,y)\,\mathrm{d}y\Big)\mathrm{d}z.$$

(Proof: calculate $\mathbb{P}(X + Y \leqslant a)$ using (1.16 a).

b) If $X \perp\!\!\!\perp Y$, $X \sim f(x)\mathrm{d}x$, $Y \sim g(x)\,\mathrm{d}x$, then

$$Z = X + Y \sim \Big(\int_{-\infty}^{\infty} f(z-y)g(y)\,\mathrm{d}y\Big)\mathrm{d}z.$$

We say that the density of $X+Y$ is the *convolution* $f * g$, with

$$f * g(z) := \int_{-\infty}^{\infty} f(z-y)g(y)\,\mathrm{d}y,$$

of the densities of X and Y. Note that even though the formula does not look symmetric in f and g, we do have $f * g = g * f$. This can be seen by a change of variable in the defining integral.

(1.23) Limsup and Liminf

Let $A_n \subset \Omega$ for all n.

a) $\limsup_{n\to\infty} A_n := \lim_{m\to\infty} \bigcup_{n=m}^{\infty} A_n := \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \{\omega : \omega \in A_n$ for infinitely many $n\}$.

b) $\liminf_{n\to\infty} A_n := \lim_{m\to\infty} \bigcap_{n=m}^{\infty} A_n := \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n = \{\omega : \omega \notin A_n$ for only finitely many $n\}$.

c) We have that $\liminf A_n \subset \limsup A_n$, and that $1_{\limsup A_n}(\omega) = \limsup 1_{A_n}(\omega)$ for all ω, where the last limsup is just the definition from real analysis. Same for liminf.

d) Sometimes we write

$$\mathbb{P}(\limsup A_n) = \mathbb{P}(\{\omega : \omega \in A_n \text{ infinitely often}\} = \mathbb{P}(A_n \text{ infinitely often}) = \mathbb{P}(A_n \text{ i.o.}).$$

(1.24) First Borel-Cantelli Lemma

Let $A_n \in \mathcal{F}$ for all n.

$$\text{If } \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \quad \text{then} \quad \mathbb{P}(A_n \text{ i.o.}) = 0.$$

(1.25) Second Borel-Cantelli Lemma

Let $A_n \in \mathcal{F}$ be an independent family of sets, Then

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \quad \text{implies} \quad \mathbb{P}(A_n \text{ i.o.}) = 1.$$

(1.26) Weak laws of large numbers

Let (X_n) be a collection of real-valued RVs. Let $\bar{S}_n := \frac{1}{n}\sum_{i=1}^n X_i$.

a) If the X_i are independent and identically distributed (i.e. $\mathbb{P}_{X_i} = \mathbb{P}_{X_j}$ for all i, j), and if $\mathbb{E}(X_i^2) < \infty$, then for all $\delta > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{S}_n - \mathbb{E}(X_1)| > \delta) = 0.$$

b) Stronger version: if all the X_i are uncorrelated, and if their variances are uniformly bounded, i.e. $v := \sup_i \mathbb{V}(X_i) < \infty$, then for all positive sequences (ε_n) and all $n \in \mathbb{N}$ we have

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}(X_i))\Big| \geqslant \varepsilon_n\Big) \leqslant \frac{v}{n\varepsilon_n^2}.$$

For suitable ε_n this can be turned into a weak law of large numbers.

c) L^1-version: Let the X_i be pairwise independent, and identically distributed. Assume that $\mathbb{E}(|X_i|) < \infty$. Then for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{S}_n - \mathbb{E}(X_1)| > \delta) = 0.$$

We also did a strong law of large numbers and a central limit theorem in the last semester, but we will do them again, slightly better and/or differently. So they do not appear in this revision. The last thing is an overview over the types of convergence for random variables.

(1.27) Types of convergence for random variables

Let (X_n) be a sequence of $\mathbb{R}^d$-valued random variables. We say that

a) $X_n \to X$ **almost surely (a.s.)** if $\mathbb{P}(\limsup_{n\to\infty} |X_n - X| > 0) = 0$.
We write $X_n \to X$ a.s. in this case.

b) $X_n \to X$ **in probability** if for all $\delta > 0$, $\limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \delta) = 0$.
We write $X_n \xrightarrow{p} X$ in this case.

c) $X_n \to X$ **in** L^p if $\mathbb{E}(|X_n - X|^p) \to 0$.
We write $X_n \xrightarrow{L^p} X$ in this case.

d) $X_n \to X$ **in distribution** if the image measures $\mathbb{P}_{X_n}$ converge weakly to $\mathbb{P}_X$, i.e. if

$$\int f(x)\mathbb{P}_{X_n}(\mathrm{d}x) \to \int f(x)\mathbb{P}_X(\mathrm{d}x) \quad \text{for all bounded and continuous functions } f.$$

In probabilistic notation, the last condition reads

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X)) \text{ for all bounded and continuous functions } f.$$

We write $X_n \Rightarrow X$, or $\mathbb{P}_{X_n} \Rightarrow \mathbb{P}_X$ in this case. Convergence in distribution is the only type of convergence where the X_n do not need to be defined on the same probability space.

We end this section by giving all the relations between the different types of convergence. Some of them have been done in the last semester, but we will prove all of them now.

(1.28) Theorem

Let (X_n) and X be $\mathbb{R}^d$-valued random variables.

a) Assume $X_n \to X$ in probability. Then
(i): $X_n \Rightarrow X$;
(ii): there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ that converges to X almost surely.

b) Assume that $X_n \Rightarrow X$, and $\mathbb{P}(X = c) = 1$ for some $c \in \mathbb{R}^d$ (in other words, the image measure of X is a Dirac measure). Then
(i): $X_n \to X$ in probability.

c) Assume that $X_n \to X$ a.s. Then
(i): $X_n \to X$ in probability.
(ii): $X_n \Rightarrow X$.
(iii): if in addition $\mathbb{E}(|X|^p) \to \mathbb{E}(|X|)$, then $X_n \to X$ in L^p.

d) Assume $X_n \to X$ in L^p. Then
(i): $X_n \to X$ in probability.
(ii): $X_n \Rightarrow X$.
(iii): there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ that converges to X almost surely.
(iv): $X_n \to X$ in L^q for all $1 \leqslant q \leqslant p$.

Proof: a) we start with part (ii). By assumption, we can pick an increasing sequence of integers (n_k) with the property that $\mathbb{P}(|X_{n_k} - X| > 1/k) < 1/k^2$. We set $A_k := \{|X_{n_k} - X| > 1/k\}$ and find that $\sum_{k=1}^\infty \mathbb{P}(A_k) < \infty$. Therefore by Borel-Cantelli, $\mathbb{P}(A_k \text{ i.o.}) = 0$, and thus

$$\mathbb{P}(\lim_{k\to\infty} |X_{n_k} - X| = 0) \geqslant \mathbb{P}(|X_{n_k} - X| > 1/k \text{ only finitely often}) = 1 - \mathbb{P}(A_k \text{ i.o.}) = 1.$$

This proves (ii). For (i), assume that $X_n \not\Rightarrow X$. Then there exists $f \in C_b$, $\delta > 0$ and a sequence $(n_j)_{j\in\mathbb{N}}$ so that

$$(*) \qquad |\mathbb{E}(f(X_{n_j})) - \mathbb{E}(f(X))| > \delta \qquad \forall j.$$

Since however $X_{n_j} \to X$ in probability as $j \to \infty$, we have just proved the existence of a subsequence $X_{n_{j_k}}$ of X_{n_j} that converges to X almost surely. Since f is continuous, also $f(X_{n_{j_k}}) \to f(X)$ almost surely, and since f is bounded, dominated convergence now implies $|\mathbb{E}(f(X_{n_{j_k}})) - \mathbb{E}(f(X))| \xrightarrow{k\to\infty} 0$. This is in contradiction to $(*)$, so we conclude that $X_n \Rightarrow X$.

b) The function

$$g_\varepsilon : \mathbb{R}^d \to \mathbb{R}_0^+, \quad x \mapsto \begin{cases} 1 & \text{if } |x-c| \geqslant \varepsilon \\ \frac{|x-c|}{\varepsilon} & \text{if } |x-c| < \varepsilon \end{cases}$$

is continuous and bounded, and therefore

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \varepsilon) \leqslant \lim_{n\to\infty} \mathbb{E}(g_\varepsilon(X_n)) = 0.$$

c) We start with (i). Put $A_n := \{|X_n - X| \leqslant \varepsilon\}$. Then

$$\liminf_{n\to\infty} A_n = \{\exists m \in \mathbb{N} : \forall n \geqslant m : |X_n - X| \leqslant \varepsilon\} = \{\limsup_{n\to\infty} |X_n - X| \leqslant \varepsilon\}.$$

So by assumption, $\mathbb{P}(\liminf A_n) = 1$, and with $Y_n = 1_{A_n}$, therefore $\mathbb{E}(\liminf Y_n) = 1$. Therefore,

$$\limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 1 - \liminf_{n\to\infty} \mathbb{P}(|X_n - X| \leqslant \varepsilon) = 1 - \liminf_{n\to\infty} \mathbb{E}(Y_n) \overset{\text{Fatou}}{\leqslant} 1 - \mathbb{E}(\liminf_{n\to\infty} Y_n) = 0.$$

So $X_n \to X$ in probability. By a), then also $X_n \Rightarrow X$.

For (iii), note that $|X_n - X|^p \leqslant 2^p(|X_n|^p + |X|^p)$, and so $Y_n := 2^p(|X_n|^p + |X|^p) - |X_n - X|^p \geqslant 0$. Also, $\liminf_{n\to\infty} Y_n = 2^{p+1}|X|^p$ almost surely. Thus,

$$2^{p+1}\mathbb{E}(|X^p|) - \limsup_{n\to\infty} \mathbb{E}(|X_n - X|^p) = \liminf \mathbb{E}(Y_n) \overset{\text{Fatou}}{\geqslant} \mathbb{E}(\liminf_{n\to\infty} Y_n) = 2^{p+1}\mathbb{E}(|X|^p),$$

and the claim follows.

d) (i) is shown by

$$\mathbb{P}(|X_n - X| \geqslant \varepsilon) \overset{\text{Chebyshev}}{\leqslant} \frac{1}{\varepsilon^p}\mathbb{E}(|X_n - X|^p) \xrightarrow{n\to\infty} 0.$$

(ii) now follows from a(i), and (iii) from a(ii). For (iv), let $q < p$, and set $\tilde{p} = \frac{p}{q}$, $\tilde{q} = \frac{p}{p-q}$. Then $1/\tilde{p} + 1/\tilde{q} = 1$, and Hölders inequality gives for all random variables Y that

$$\|Y\|_q^q = \mathbb{E}(|Y|^q \cdot 1) \leqslant \mathbb{E}(|Y|^{q\tilde{p}})^{1/\tilde{p}} \mathbb{E}(1^{\tilde{q}})^{1/\tilde{q}} = \mathbb{E}(|Y|^p)^{q/p} = \|Y\|_p^q,$$

so $\|Y\|_q \leqslant \|Y\|_p$ if $q \leqslant p$. Applying this to $Y = |X_n - X|$ shows the claim. □

(1.29) Diagram

The following diagram summarizes all the relations of the previous theorem.

$$\begin{tikzcd}
X_n \to X \text{ a.s.} \arrow[r, bend left, "\text{if } \mathbb{E}(|X_n|^p) \to \mathbb{E}(|X|^p)"] \arrow[d, bend right, "\text{always}"'] \arrow[dr, "\text{always}"] & X_n \to X \text{ in } L^p \arrow[l, bend left, "\text{along a subsequence}"] \arrow[dl, "\text{always}"'] \arrow[d, "\text{always}"] \\
X_n \to X \text{ in probability} \arrow[u, bend right, "\text{along a subsequence}"'] \arrow[r, bend right, "\text{always}"'] & X_n \Rightarrow X \arrow[l, bend right, "\text{if } \mathbb{P}_X \text{ is a Dirac measure}"']
\end{tikzcd}$$

(1.30) Warning

Note that since $x \mapsto x$ and $x \mapsto x^2$ are not bounded functions, convergence in distribution does *not* imply that $\mathbb{E}(X_n) \to \mathbb{E}(X)$ or $\mathbb{V}(X_n) \to \mathbb{V}(X)$. To assume this is a very tempting and common mistake.

2. Sums of independent random variables

We are now interested in what happens when we take infinitely many independent random variables and sum them up. We start by making sure that there exist a probability space that is large enough to define all those independent random variables.

Part 1: Existence of countably many independent random variabes

(2.1) Infinite products of measurable spaces

a) Let $(\Omega_i, \mathcal{G}_i)$, $i \in \mathbb{N}$, be measurable spaces. The set

$$\Omega := \Omega_1 \times \Omega_2 \times \ldots := \bigoplus_{i=1}^{\infty} \Omega_i := \{(\omega_i)_{i \in \mathbb{N}} : \omega_i \in \Omega_i \, \forall i\}$$

is called the *direct sum* of the sets Ω_i. We define the σ-algebras

$$\mathcal{F}_{\{n\}} := \{\Omega_1 \times \cdots \times \Omega_{n-1} \times A \times \Omega_{n+1} \times \ldots : A \in \mathcal{G}_n\}$$

and

$$\mathcal{F}_n := \sigma\Big(\bigcup_{j=1}^{n} \mathcal{F}_{\{j\}}\Big), \quad \mathcal{F} := \sigma\Big(\bigcup_{j=1}^{\infty} \mathcal{F}_{\{j\}}\Big),$$

Recall that if $\mathcal{A}$ is a collection of sets, then $\sigma(\mathcal{A})$ is the smallest σ-algebra containing $\mathcal{A}$.

b) The σ-algebras $\mathcal{F}_n$ have the property that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$. Such a collection of σ-algebras is called a *filtration*. It will become important later in the lecture when we discuss martingales.

c) In most examples, all the $(\Omega_i, \mathcal{G}_i)$ are the same. Then we write

$$\bigoplus_{i=1}^{\infty} \Omega_i = \Omega_1^{\mathbb{N}}, \quad \mathcal{F} = \mathcal{G}_1^{\otimes \mathbb{N}}.$$

Note that even in that case, $\mathcal{F}_{\{i\}} \neq \mathcal{F}_{\{j\}}$ if $i \neq j$.

We now work towards proving the existence of independent RVs. The first step is

(2.2) Proposition

Let $\Omega = \{0,1\}^{\mathbb{N}}$, $\mathcal{F} = \mathcal{P}(\{0,1\})^{\otimes \mathbb{N}}$. There exists a probability measure μ on Ω with the property that for all $q = (q_1, \ldots, q_n) \in \{0,1\}^n$, we have

$$\mu(\{\omega \in \{0,1\}^{\mathbb{N}} : \omega_i = q_i \, \forall i \leqslant n\}) = 2^{-n}.$$

Proof: Consider the probability space $((0,1], \mathcal{B}((0,1]), \lambda)$ where λ is the Lebesgue measure. For $x \in (0,1]$ consider its unique non-terminating dyadic representation:

$$x = \sum_{i=1}^{\infty} q_i 2^{-i} \qquad \text{with } q_i \in \{0,1\}, \limsup_{i \to \infty} q_i = 1.$$

(e.g. 0.1 is represented with $q_1 = 0$ and $q_j = 1$ if $j > 1$.) Now define the random variable

$$f : (0,1] \to \{0,1\}^{\mathbb{N}}, \quad x \mapsto (q_i)_{i \in \mathbb{N}}.$$

f is indeed measurable because

$$f^{-1}\big(\{q_1\} \times \ldots \{q_n\} \times \{0,1\}^{\mathbb{N}}\big) = \big(\sum_{i=1}^{n} q_i 2^{-i}, \sum_{i=1}^{n} q_i 2^{-i} + 2^{-n}\big] \in \mathcal{B},$$

and the sets $\{q_1\} \times \ldots \{q_n\} \times \{0,1\}^{\mathbb{N}}$ generate $\mathcal{F}_n$ for each n, and therefore generate $\mathcal{F}$ when we consider all n. Now we can use (1.11). Since $\mu = \lambda \circ f^{-1}$, the claim is shown. □

(2.3) Corollary

A sequence of independent Bernoulli(1/2) random variables exists.

Proof: Take Ω as in (2.2) and $X_i(\omega) = \omega_i$. Then for each finite set $i_1, \ldots, i_n$ of different integers, and all $q_1, \dots, q_n \in \{0,1\}$, simple combinatorics give

$$\mu(X_{i_1} = q_1, \dots X_{i_n} = q_n) = \mu\big(\{\omega \in \{0,1\}^{\mathbb{N}} : \omega_i = q_i \forall i \leqslant n\}\big) = 2^{-n} = \prod_{i=1}^{n} \mu(X_{i_j} = q_j).$$

This shows independence and $B(1/2)$-distribution. □

The following Lemma goes the opposite way as (2.2):

(2.4) Lemma

Let μ be defined as in (2.2), and let

$$g : \{0,1\}^{\mathbb{N}} \to [0,1], \quad (q_i)_{i\in\mathbb{N}} \mapsto \sum_{i=1}^{\infty} q_i 2^{-i}.$$

Then $\mu \circ g^{-1} = \lambda_{[0,1]}$.

Proof: g is 'almost' equal to f^{-1} with f from the proof of (2.2). The only difference is for sequences ending in $0, 0, \dots$. We have

$$g^{-1}\Big(\big[\sum_{i=1}^{n} q_i 2^{-i}, \sum_{i=1}^{n} q_i 2^{-i} + 2^{-n}\big]\Big) = \{q_1\} \times \{q_2\} \times \cdots \times \{q_n\} \times \{0,1\}^{\mathbb{N}} \in \mathcal{F}.$$

Since the intervals generate $\mathcal{B}$, g is measurable, and

$$\mu \circ g^{-1}\Big(\big[\sum_{i=1}^{n} q_i 2^{-i}, \sum_{i=1}^{n} q_i 2^{-i} + 2^{-n}\big]\Big) = 2^{-n} = \lambda\Big(\big[\sum_{i=1}^{n} q_i 2^{-i}, \sum_{i=1}^{n} q_i 2^{-i} + 2^{-n}\big]\Big).$$

Since the closed intervals form a $\cap$-stable generator of $\mathcal{B}$, the π-λ-Theorem implies that

$$\mu \circ g^{-1}(A) = \lambda(A) \quad \text{for all } A \in \mathcal{B}.$$

□

The last Lemma can be used to prove the existence of infinitely many real-valued RVs:

(2.5) Proposition

a) Let $\Omega = [0,1]^{\mathbb{N}}$, $\mathcal{F} = \mathcal{B}([0,1])^{\otimes\mathbb{N}}$. There exists a measure $\lambda_{[0,1]}^{\otimes\mathbb{N}}$ on Ω with

$$(*) \qquad \lambda_{[0,1]}^{\otimes\mathbb{N}}\Big([a_1, b_1) \times \cdots \times [a_n, b_n) \times [0,1]^{\mathbb{N}}\Big) = \prod_{i=1}^{n} (b_i - a_i)$$

for all $0 \leqslant a_i < b_i \leqslant 1$. Consequently, there exists a sequence of independent $\mathcal{U}[0,1]$-distributed (i.e. uniformly distributed on $[0,1]$) RVs.

b)Let $\Omega = \mathbb{R}^{\mathbb{N}}$, $\mathcal{F} = \mathcal{B}^{\otimes\mathbb{N}}$. Let μ_i, $i \in \mathbb{N}$ be probability measures on $\mathbb{R}$. Then the product measure $\bigotimes_{i=1}^{\infty} \mu_i$, i.e. the unique measure with

$$\bigotimes_{i=1}^{\infty} \mu_i\Big([a_1, b_1) \times \cdots \times [a_n, b_n) \times \mathbb{R}^{\mathbb{N}}\Big) = \prod_{i=1}^{n} \mu_i([a_i, b_i)) \quad \forall a_i \leqslant b_i$$

exists. Consequently, a sequence of independent RVs X_i with $X_i \sim \mu_i$ exists.

Proof: Let $p_1, p_2, \dots$ be the sequence of prime numbers, and consider the map

$$G : \{0,1\}^{\mathbb{N}} \to [0,1]^{\mathbb{N}}, \quad (q_i)_{i\in\mathbb{N}} \mapsto \big(\sum_{j=1}^{\infty} q_{p_i^j} 2^{-j} \big)_{i\in\mathbb{N}} = \big(0.q_2 q_4 q_8 q_{16} \cdots, 0.q_3 q_9 q_{27} \cdots, 0.q_5 q_{25} q_{125} \cdots, \dots \big),$$

where in the last expression we used binary digits. You should check as an exercise that G is measurable with respect to the product σ-algebras. Since no q_j is used twice (many are not used at all!), each component of G has the same distribution as g from (2.4), and the components are alle independent. This proves a).

For b), consider the distribution function transform: Let F_i be the distribution function of μ_i, and put

$$f_i(x) := \sup\{z \in \mathbb{R} : F_i(z) < x\}.$$

Then you can (and should) check that $\mu_i = \lambda_{[0,1]} \circ f_i^{-1}$. So, we can apply this transform to each coordinate of the measure in a) and prove b). □

Nomenclature We will write iid for 'independent, identically distributed'.

Part 2: Strong laws of large numbers

We have already shown the following strong law in the previous course:

(2.6) Strong law under fourth moment condition

Let (X_n) be a sequence of iid random variables, and assume $\mathbb{E}(X_i) = \mu$ and $\mathbb{E}(X_i^4) < \infty$ for all i. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \qquad \text{almost surely.}$$

The next theorem (the main result of this subsection) states the same convergence under weaker conditions.

(2.7) Theorem: Strong law of large numbers

Let (X_i) be pairwise independent and identically distributed RVs, with $\mathbb{E}(|X_i|) < \infty$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \qquad \text{almost surely.}$$

Proof: We do the proof in several steps.

Step 1: Truncation. Let $Y_n(\omega) := X_n(\omega) 1_{\{|X_n| \leqslant n\}}$, and

$$S_n := \sum_{i=1}^{n} X_i, \quad T_n := \sum_{i=1}^{n} Y_i.$$

We claim that

$$\lim_{n\to\infty} \tfrac{1}{n} T_n = \mu \text{ a.s.} \quad \text{implies} \qquad \lim_{n\to\infty} \tfrac{1}{n} S_n = \mu \text{ a.s.}$$

Indeed, we have

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_k| > k) = \sum_{k=1}^{\infty} \mathbb{P}(|X_1| > k) \leqslant \int_0^{\infty} \mathbb{P}(|X_1| > t)\,\mathrm{d}t = \mathbb{E}(|X_1|) < \infty,$$

and so the first Borel-Cantelli Lemma implies that with $A := \{Y_n \neq X_n \text{ i.o.}\}$, we have $\mathbb{P}(A) = 0$. It follows that for all $\omega \notin A$, there exists $N(\omega) < \infty$ with

$$\sup_{n\in\mathbb{N}} |S_n(\omega) - T_n(\omega)| \leqslant \sum_{k=1}^{N(\omega)} X_k(\omega) 1_{\{|X_k(\omega)|>k\}} < \infty.$$

This implies that

$$\lim_{n\to\infty} \left|\tfrac{1}{n} T_n(\omega) - \tfrac{1}{n} S_n(\omega)\right| = 0$$

for all $\omega \notin A$ and finishes step 1.

Step 2: We prove the estimate

$$\sum_{k=1}^{\infty} \frac{\mathbb{V}(Y_k)}{k^2} \leqslant 4\mathbb{E}(|X_1|) < \infty,$$

which we will need shortly. We start by computing

$$\mathbb{V}(Y_k) \leqslant \mathbb{E}(Y_k^2) \overset{(1.16e)}{=} \int_0^{\infty} 2y\mathbb{P}(|Y_k| > y)\,\mathrm{d}y \leqslant \int_0^{\infty} 2y\mathbb{P}(|X_1| > y) 1_{\{y\,\leqslant\,k\}}\,\mathrm{d}y.$$

So summing over k and using Fubini (everything is nonnegative!) gives

$$(*) \quad \sum_{k=1}^{\infty} \frac{\mathbb{V}(Y_k)}{k^2} \leqslant \sum_{k=1}^{\infty} \frac{1}{k^2} \int_0^{\infty} 2y\mathbb{P}(|X_1| > y) 1_{\{y\,\leqslant\,k\}}\,\mathrm{d}y = 2\int_0^{\infty} \Big(y \sum_{k=1}^{\infty} \frac{1}{k^2} 1_{\{y<k\}}\Big) \mathbb{P}(|X_1| > y)\,\mathrm{d}y.$$

A little analysis gives that for $y \geqslant 1$,

$$y \sum_{k=1}^{\infty} \frac{1}{k^2} 1_{\{y<k\}} \leqslant y \int_{\lfloor y \rfloor}^{\infty} \frac{1}{x^2}\,\mathrm{d}x = \frac{y}{\lfloor y \rfloor} \leqslant 2,$$

and for $0 \leqslant y \leqslant 1$,

$$y \sum_{k=1}^{\infty} \frac{1}{k^2} 1_{\{y<k\}} \leqslant \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} < 2.$$

Using this and (1.16 e) with $p = 1$ in $(*)$ shows step 2.

Step 3: We now prove the claim for the T_n, but for the moment only along a subsequence. Let $k(n) = \lfloor \alpha^n \rfloor$ with some $\alpha > 1$. Then by the Chebyshev inequality, for each $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\big(|T_{k(n)} - \mathbb{E}(T_{k(n)})| > \varepsilon k(n)\big) \leqslant \sum_{n=1}^{\infty} \frac{\mathbb{V}(T_{k(n)})}{\varepsilon^2 k(n)^2} =$$

$$= \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{k(n)^2} \sum_{m=1}^{k(n)} \mathbb{V}(Y_m) = \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \mathbb{V}(Y_m) \sum_{n:k(n)\,\geqslant\,m} \frac{1}{k(n)^2}.$$

In the last step, we re-ordered the sum of nonnegative terms. Since $\alpha^n \geqslant k(n) \geqslant \alpha^n/2$, we have

$$\sum_{n:k(n)\geqslant m} \frac{1}{k(n)^2} \leqslant 4 \sum_{n:\alpha^n \geqslant m} \alpha^{-2n} \leqslant \frac{4}{m^2} \sum_{n=0}^{\infty} \alpha^{-2n} = \frac{4}{m^2(1-\alpha^{-2})}.$$

We conclude that, by step 2 and our calculations,

$$\sum_{n=1}^{\infty} \mathbb{P}\big(|T_{k(n)} - \mathbb{E}(T_{k(n)})| > \varepsilon k(n)\big) \leqslant \frac{4}{\varepsilon^2(1-\alpha^{-2})} \sum_{m=1}^{\infty} \frac{\mathbb{V}(Y_m)}{m^2} \leqslant \frac{16}{\varepsilon^2(1-\alpha^{-2})}\mathbb{E}(|X_1|) < \infty.$$

By Borel-Cantelli, this means that for each $\varepsilon > 0$,

$$\mathbb{P}\big(\frac{1}{k(n)}|T_{k(n)} - \mathbb{E}(T_{k(n)})| > \varepsilon \text{ i.o.}\big) = 0,$$

and so in particular with

$$A_m := \big\{\omega \in \Omega : \limsup_{n\to\infty} \frac{1}{k(n)}|T_{k(n)} - \mathbb{E}(T_{k(n)})| > \frac{1}{m}\big\},$$

we find $\mathbb{P}(\bigcap_m A_m^c) = \mathbb{P}((\bigcup_m A_m)^c) = 1 - \mathbb{P}(\bigcup_m A_m) \geqslant 1 - \sum_m \mathbb{P}(A_m) = 1$. So,

$$\bigcap_m A_m^c = \big\{\forall m : \limsup_{n\to\infty} \frac{1}{k(n)}|T_{k(n)} - \mathbb{E}(T_{k(n)})| \leqslant \frac{1}{m}\big\} = \{\limsup_{n\to\infty} \frac{1}{k(n)}|T_{k(n)} - \mathbb{E}(T_{k(n)})| = 0\}.$$

Finally, $\lim_{k\to\infty} \mathbb{E}(Y_k) = \mathbb{E}(X_1)$, and for any sequence (a_n) such that $a_n \to a$ we have that $\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^N a_n = a$ (exercise!). Applying this to the sequence $\mathbb{E}(Y_k)$ gives

$$\lim_{n\to\infty} \frac{1}{k(n)}\mathbb{E}(T_{(k(n))}) = \lim_{n\to\infty} \frac{1}{k(n)} \sum_{j=1}^{k(n)} \mathbb{E}(Y_j) = \mathbb{E}(X_1).$$

We have finished step 3.

Step 4: We now prove the full claim, but only for nonnegative RVs. If $X_i(\omega) \geqslant 0$ for all i and all ω, then also $Y_i(\omega) \geqslant 0$ for all i, ω. Therefore with $k(n)$ as in step 3,

$$\frac{T_{k(n)}(\omega)}{k(n+1)} \leqslant \frac{T_m(\omega)}{m} \leqslant \frac{T_{k(n+1)}(\omega)}{k(n)}$$

for all ω, all n, and all $m \in [k(n), k(n+1)]$. Since $k(n+1)/k(n) \to \alpha$ as $n \to \infty$, we find a.s.

$$\frac{1}{\alpha}\mathbb{E}(X_1) \leqslant \liminf_{m\to\infty} \tfrac{T_m}{m} \leqslant \limsup_{m\to\infty} \tfrac{T_m}{m} \leqslant \alpha\mathbb{E}(X_1).$$

Since $\alpha > 1$ was arbitrary, the claim holds for all nonnegative (X_n).

Step 5: For general (X_n), we just decompose into positive and negative part: $X_n = X_{n,+} - X_{n,-}$. By step 4, the claim holds for $X_{n,\pm}$ and so also for X_n. □

The strong law also holds in the situation where $\mathbb{E}(X_1) = +\infty$:

(2.8) Theorem

Let (X_i) be iid with $\mathbb{E}(X_{1,+}) = \infty$, $\mathbb{E}(X_{1,-}) < \infty$. Then with $S_n = \sum_{i=1}^n X_i$, we have $\lim \frac{1}{n}S_n = +\infty$ almost surely.

Proof: Let $X_i^M(\omega) = \min\{X_i(\omega), M\}$, and $S_n^M = \sum_{i=1}^n X_i^M$. Then almost surely,

$$\liminf_{n\to\infty} \frac{S_n}{n} \geqslant \lim_{n\to\infty} \frac{1}{n} S_n^M \overset{(2.7)}{=} \mathbb{E}(X_1^M).$$

By monotone convergence, $\mathbb{E}(X_{1,+}^M) \to \infty$ as $M \to \infty$, and since $\mathbb{E}(X_{1,-}^M) = \mathbb{E}(X_{1,-}) < \infty$, the claim follows. □

Next we give a few applications for the strong law of large numbers.

(2.9) Example: Renewal Theory

Let (X_i) be iid RVs, with $0 < X_i < \infty$. We interpret X_i as the waiting time between two events, e.g. the time it takes between two clicks of a Geiger counter. Then

$$T_n(\omega) := \sum_{i=1}^n X_i(\omega)$$

is the total time it takes before we see the n-th event, and

$$N_t(\omega) := \sup\{n \in \mathbb{N} : T_n(\omega) \leqslant t\}$$

is the total number of events up to (and including) time t. We want to understand how N_t behaves for large t.

(2.10) Theorem

In the situation of the example above, write $\mu := \mathbb{E}(X_1) \leqslant \infty$. Then

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu} \quad \text{a.s.}$$

Proof: By definition of N_t, we have for all $\omega \in \Omega$

$$T_{N_t(\omega)}(\omega) = \sum_{i=1}^{N_t(\omega)} X_i(\omega) \leqslant t < T_{N_t(\omega)+1}(\omega),$$

and thus

$$(*) \qquad \frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \leqslant \frac{t}{N_t(\omega)} < \frac{T_{N_t(\omega)+1}(\omega)}{N_t(\omega)+1} \frac{N_t(\omega)+1}{N_t(\omega)}.$$

Since $X_i(\omega) < \infty$ for all ω, we have that $\lim_{t\to\infty} N_t(\omega) = \infty$ for all ω. By (2.7) or (2.8), there is a set Ω_0 with $\mathbb{P}(\Omega_0) = 1$ and $\frac{T_n(\omega)}{n} \to \mu$ for all $\omega \in \Omega_0$. Then on Ω_0, the left hand side and right hand side of $(*)$ both converge to μ. The claim follows. □

(2.11) Remark and Exercise

In the previous proof, we used the fact that if $X_n \to X$ a.s., and if $N_n \to \infty$ a.s., then also $X_{N_n} \to X$ a.s. This becomes false when we replace the a.s. convergence of X_n by convergence in probability. Exercise: give an example of a sequence (X_n) with $X_n \in \{0,1\}$, $X_n \to 0$ in probability, $N(n) \to \infty$ a.s. and $X_{N(n)} \to 1$ a.s..

(2.12) Empirical measure and empirical distribution function

Let (X_i) be real random variables. We write δ_x for the Dirac measure at $x \in \mathbb{R}$. The map $\Omega \to$ (probability measures on $\mathbb{R}$),

$$\omega \mapsto \mu_{n,\omega} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

(i.e.: the random probability measure μ_n) is called the *empirical measure* of the random variables $X_1, \dots, X_n$. $\mu_{n,\omega}$ is the normalized histogram of values that we have obtained by observing the first n of the X_i for the choice of randomness ω.

The distribution function F_n of μ_n, i.e. the random function $\omega \mapsto F_{n,\omega}$ with

$$F_{n,\omega}(x) = \mu_{n,\omega}((-\infty, x]) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i(\omega) \leqslant x\}}$$

is called the *empirical distribution function* of the X_i. $F_{n,\omega}(x)$ is the fraction of values $\leqslant x$ that have been observed after n observations of the X_i, for the particular incarnation ω of the randomness. If the X_i are iid, then the RVs $Y_i = 1_{\{X_i \leqslant x\}}$ are iid and integrable. Therefore by (2.7),

$$F_{n,\omega}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \overset{n\to\infty}{\longrightarrow} \mathbb{E}(Y_1) = \mathbb{P}(X_1 \leqslant x) = F(x),$$

almost surely. The following result shows that convergence is even uniform in x.

(2.13) Glivenko-Cantelli theorem

In (2.12), assume that the X_i are iid, and let F be the distribution function for X_1. Then

$$\lim_{n\to\infty} \|F_n - F\|_\infty = \lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0 \quad \text{a.s.}$$

Proof: Pointwise convergence has been shown in (2.12). By the same argument as there, we can set $Z_n = 1_{\{X_n < x\}}$ and find

$$F_n(x-) := \lim_{y \nearrow x} F_n(y) = \frac{1}{n} \sum_{i=1}^{n} Z_i \overset{n\to\infty}{\longrightarrow} \mathbb{E}(Z_1) = \lim_{y \nearrow x} F(y) =: F(x-)$$

almost surely, for each fixed x. Now choose $k \in \mathbb{N}$, then there is a set Ω_k with $\mathbb{P}(\Omega_k) = 1$ and, for each $\omega \in \Omega_k$ a number $N(\omega) \in \mathbb{N}$ so that with $x_{j,k} = \inf\{y \in \mathbb{R} : F(y) \geqslant j/k\}$, j = 1, …, k-1, $x_{0,k} = -\infty$, $x_{k,k} = \infty$, we have

$$\max\{|F_{n,\omega}(x_{j,k}) - F(x_{j,k})| : 0 \leqslant j \leqslant k, n \geqslant N(\omega)\} \leqslant \frac{1}{k}$$

and

$$\max\{|F_{n,\omega}(x_{j,k}-) - F(x_{j,k}-)| : 0 \leqslant j \leqslant k, n \geqslant N(\omega)\} \leqslant \frac{1}{k}.$$

Let $x \in (x_{j-1,k}, x_{j,k})$ for some $0 \leqslant j \leqslant k$. Since F is monotone and $F(x_{j,k}-) - F(x_{j-1,k}) \leqslant \frac{1}{k}$, we find that for all $\omega \in \Omega_k$, all $x \in \mathbb{R}$ and all $n \geqslant N(\omega)$:

$$F_{n,\omega}(x) \leqslant F_{n,\omega}(x_{j,k}-) \leqslant F(x_{j,k}-) + \tfrac{1}{k} \leqslant F(x_{j-1,k}) + \tfrac{2}{k} \leqslant F(x) + \tfrac{2}{k},$$

and

$$F_{n,\omega}(x) \geqslant F_{n,\omega}(x_{j-1,k}) \geqslant F(x_{j-1,k}) - \tfrac{1}{k} \geqslant F(x_{j,k}-) - \tfrac{2}{k} \geqslant F(x) - \tfrac{2}{k}.$$

We conclude that $\limsup_{n\to\infty} \|F_{n,\omega} - F\|_\infty \leqslant \frac{2}{k}$ for all $\omega \in \Omega_k$. Consequently, for all $\omega \in \Omega_\infty := \bigcap_{k\in\mathbb{N}} \Omega_k$, we find that $\lim_{n\to\infty} \|F_{n,\omega} - F\|_\infty = 0$. Since $\mathbb{P}(\Omega_\infty) = 1$, the result follows. □

Part 3: Large Deviations

(2.14) Motivation

Let (X_i) be iid integrable RVs, $S_n = \sum_{i=1}^n X_i$. By the weak law of large numbers,

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{1}{n}S_n - \mathbb{E}(X_1)| \geqslant a) = 0$$

for all $a > 0$. We are now interested in the rate of convergence, i.e. how quickly $\mathbb{P}(|\frac{1}{n}S_n - \mathbb{E}(X_1)| \geqslant a)$ decays to zero. Since in applications, we always have large but finite n, such an information is often more valuable than just the convergence.

We will find that under mild conditions, the decay is exponentially fast, and we can give an (usually implicit) formula for the rate of exponential decay. We concentrate on the case $\mathbb{E}(X_1) = 0$ and the expression $\mathbb{P}(\frac{1}{n}S_n \geqslant a)$. The case of $\mathbb{E}(X_1) \neq 0$ and $\mathbb{P}(\frac{1}{n}S_n \leqslant -a)$ can then be easily deduced from our results.

(2.15) Existence of exponential rate

Let (X_i) be a sequence of iid random variables. Then, with $S_n = \sum_{i=1}^n X_i$, the limit

$$\gamma(a) := \lim_{n\to\infty} \tfrac{1}{n} \ln \mathbb{P}(\tfrac{1}{n}S_n \geqslant a) = \sup_{n\in\mathbb{N}} \tfrac{1}{n} \ln \mathbb{P}(\tfrac{1}{n}S_n \geqslant a)$$

exists for all $a \in \mathbb{R}$ and is in $[-\infty, 0]$. Furthermore, $\gamma(a) = -\infty$ if and only if $\mathbb{P}(X_1 \geqslant a) = 0$.

Proof: Let $\pi_n := \mathbb{P}(\frac{1}{n}S_n \geqslant a) = \mathbb{P}(S_n \geqslant na)$. If $\mathbb{P}(X_1 \geqslant a) = 0$, then $\mathbb{P}(S_n \geqslant na) = 0$ for all n and $\gamma(a) = -\infty$. Let now $\mathbb{P}(X_1 \geqslant a) > 0$. S_n and $S_{n+m} - S_n$ are independent, and thus

$$\begin{aligned}\pi_{n+m} &= \mathbb{P}(S_{n+m} \geqslant (n+m)a) \geqslant \mathbb{P}(S_n \geqslant na, S_{n+m} - S_n \geqslant ma)\\ &= \mathbb{P}(S_n \geqslant na)\mathbb{P}(S_{n+m} - S_n \geqslant ma) = \pi_n \pi_m > 0.\end{aligned}$$

So taking logarithms (ln is monotone!) we find with $\gamma_n := ln \pi_n$, $\gamma_{m+n} \geqslant \gamma_n + \gamma_m$. By the lemma below (with $\gamma_n = -x_n$), this shows that $\lim_{n\to\infty} \frac{1}{n}\gamma_n = \sup_{n\in\mathbb{N}} \frac{1}{n}\gamma_n$ exists in $(-\infty, \infty]$ and is in particular not $-\infty$. Since $\pi_n = \mathbb{P}(\frac{1}{n}S_n \geqslant a) \leqslant 1$, the limit is $\leqslant 0$. □

The next lemma is needed in the proof above. It is so frequently useful that it deserves its own number.

(2.16) Subadditive Lemma, or Fetekes Lemma

Let (x_n) be a sequence that is *subadditive*, i.e. such that

$$x_{n+m} \leqslant x_n + x_m \qquad \forall n, m \in \mathbb{N}.$$

Then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{n\in\mathbb{N}} \frac{x_n}{n}$$

exists in $[-\infty, \infty)$.

Proof: Fix $m \in \mathbb{N}$, and write $n > m$ in the form $n = km + l$, $k \in \mathbb{N}$, with $0 \leqslant l < m$. Then

$$x_n \leqslant x_{km} + x_l \leqslant x_{(k-1)m} + x_m + x_l = \ldots = kx_m + x_l.$$

Thus

$$\frac{x_n}{n} \leqslant \frac{kx_m}{n} + \frac{x_l}{n} = \frac{km}{km+l}\frac{x_m}{m} + \frac{x_l}{n} \leqslant \frac{x_m}{m} + \frac{x_l}{n}.$$

Taking $\limsup_{n\to\infty}$ and using that $l = l(n) < m$, we find that

$$\limsup_{n\to\infty} \frac{x_n}{n} \leqslant \frac{x_m}{m}$$

for all $m \in \mathbb{N}$. Taking now $\inf_{m\in\mathbb{N}}$ and $\liminf_{m\to\infty}$ in the equation above, and using that $\frac{x_1}{1} < \infty$, shows all the claims. □

(2.17) General large deviation estimate

Let (X_i) be iid integrable RVs, $\mathbb{E}(X_1) = 0$, and define $\gamma(a)$ as in (2.15). Then the function $a \mapsto \gamma(a)$ is concave on $[0, \inf\{a : \gamma(a) = -\infty\})$, and for all $n \in \mathbb{N}$,

$$(*) \qquad \mathbb{P}(\tfrac{1}{n}S_n \geqslant a) \leqslant \mathrm{e}^{n\gamma(a)} \;=\; \mathrm{e}^{-n|\gamma(a)|}.$$

Also, $\gamma(a) = 0$ if $a < \mathbb{E}(X_1) = 0$ and $\gamma(a) = -\infty$ if $\mathbb{P}(X_1 \geqslant a) = 0$.

Proof: Equation $(*)$ follows from (2.15):

$$\frac{1}{n}\ln \mathbb{P}(\tfrac{1}{n}S_n \geqslant a) \leqslant \sup_m \frac{1}{m}\ln \mathbb{P}(\tfrac{1}{m}S_m \geqslant a) = \gamma(a)$$

for all n. Now rearrange. The two statements after $(*)$ are also clear, the first one follows from the law of large numbers. To see the concavity, note that the same trick as in the proof of (2.15) gives

$$\mathbb{P}(S_{qm} \geqslant pma + (q-p)mb) \geqslant \mathbb{P}(S_{pm} \geqslant pma)\mathbb{P}(S_{(q-p)m} \geqslant (q-p)mb),$$

for all $p, q \in \mathbb{N}$ with $q > p$ and all $a, b \in \mathbb{R}$. With $\lambda = p/q$, this means that

$$\frac{1}{qm}\ln \mathbb{P}(\tfrac{1}{qm}S_{qm} \geqslant \lambda a + (1-\lambda)b) \geqslant \lambda \frac{1}{pm}\ln \mathbb{P}(\tfrac{1}{pm}S_{pm} \geqslant a) + (1-\lambda)\frac{1}{(q-p)m}\ln \mathbb{P}(\tfrac{1}{(q-p)m}S_{(q-p)m} \geqslant b).$$

Taking the limit $m \to \infty$ on both sides above shows

$$\gamma(\lambda a + (1-\lambda)b) \geqslant \lambda\gamma(a) + (1-\lambda)\gamma(b)$$

for all rational $\lambda \in (0,1)$. Since $a \mapsto \gamma(a)$ is monotone, it is easy to extend this to all real λ. □

The existence of some $\gamma(a)$ is nice, but more interesting is to actually compute its value; in particular, if $\gamma(a) = 0$ then (2.17) is not very useful. We calculate $\gamma(a)$ in two special cases.

(2.18) Example: Normal distributions

If $X_i \sim \mathcal{N}(0,1)$ for all i in (2.17), then $S_n \sim \mathcal{N}(0,n)$ and thus

$$\mathbb{P}(S_n \geqslant an) = \mathbb{P}(X_1 \geqslant a\sqrt{n}) = \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{n}}^{\infty} \mathrm{e}^{-x^2/2}\, \mathrm{d}x$$

An important fact about the Gaussian density (exercise!) is that

$$\frac{1}{c+1/c}\,\mathrm{e}^{-c^2/2} \leqslant \int_c^\infty \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leqslant \frac{1}{c}\,\mathrm{e}^{-c^2/2}$$

for all $c > 0$. Therefore

$$\gamma(a) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(S_n \geqslant an) = -\frac{a^2}{2},$$

for all $a > 0$.

(2.19) Example: Coin flips

If $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$ for all i in (2.17), then for $a \in [0,1]$

$$(*) \qquad \gamma(a) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(S_n \geqslant an) = -\Big(\frac{1+a}{2}\ln(1+a) + \frac{1-a}{2}\ln(1-a)\Big).$$

For $a > 1$, $\gamma(a) = -\infty$ as we know from (2.17). Note that $\gamma(1) = -\ln 2$ (convention: $0 \ln 0 = \lim_{x\searrow 0} x \ln x = 0$), so γ is 'discontinuous' at $a = 1$. To prove formula $(*)$, note first that the case $a = 0$ is clear. For $0 < a < 1$, we have

$$\mathbb{P}(S_n \geqslant an) = 2^{-n} \sum_{k=\lfloor (1+a)n/2+1 \rfloor}^{n} \binom{n}{k}.$$

Since $\binom{n}{k} \leqslant \binom{n}{\lfloor (1+a)n/2+1\rfloor}$ for $a > 0$ and all $k \geqslant \lfloor (1+a)n/2 + 1\rfloor$, this means that

$$2^{-n}\binom{n}{\lfloor (1+a)n/2+1\rfloor} \leqslant \mathbb{P}(S_n \geqslant an) \leqslant (n+1)2^{-n}\binom{n}{\lfloor (1+a)n/2+1\rfloor}.$$

This already shows that

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(S_n \geqslant an) = -\ln 2 + \lim_{n\to\infty} \frac{1}{n} \ln \binom{n}{\lfloor (1+a)n/2+1\rfloor},$$

since the term $\frac{1}{n}\ln(n+1)$ on the right hand side above disappears in the limit. The rest is a direct calculation using Stirling's formula:

$$\lim_{n\to\infty} \frac{1}{n!} n^n \,\mathrm{e}^{-n}\, \sqrt{2\pi n} = 1.$$

We now want to give an (abstract) general formula for $\gamma(a)$. We need the following definition:

(2.20) Definition

For a random variable X, the function

$$\varphi_X : \mathbb{R} \to (0,\infty], \quad \theta \mapsto \varphi_X(\theta) := \mathbb{E}(\,\mathrm{e}^{\theta X}\,)$$

is called the *moment generating function* of the random variable X.

Note that $\varphi_X(0) = 1$ always, but it is possible that $\varphi_X(\theta) = \infty$ for all $\theta \neq 0$. Our first result is an upper bound for $\gamma(a)$.

(2.21) Proposition

In the situation of (2.17), with $\varphi(\theta) := \varphi_{X_1}(\theta)$, we have

$$\gamma(a) \leqslant\, -\sup\{a\theta - \ln\varphi(\theta) : \theta \geqslant 0.\}$$

Proof: If $\varphi(\theta) = \infty$ for all $\theta \neq 0$, then the supremum above is taken at $\theta = 0$, and the statement is $\gamma(a) \leqslant 0$, which is trivially true (and useless). So assume now that $\varphi(\theta) < \infty$ for some $\theta > 0$. Then by Chebyshev's inequality,

$$\mathrm{e}^{\theta na}\,\mathbb{P}(S_n \geqslant na) \leqslant \mathbb{E}(\,\mathrm{e}^{\theta S_n}\,) \overset{\text{iid}}{=} \mathbb{E}(\,\mathrm{e}^{\theta X_1}\,)^n = \varphi(\theta)^n.$$

Thus,

$$\mathbb{P}(S_n \geqslant na) \leqslant\, \mathrm{e}^{-n(\theta a - \ln\varphi(\theta))}\,,$$

and taking $\lim_{n\to\infty}\frac{1}{n}\ln$ on both sides shows

$$-\gamma(a) \geqslant \theta a - \ln\varphi(\theta)$$

for all θ. Taking the supremum over θ on the right hand side of this inequality shows the claim. □

(2.22) Remark

Consider $\theta_+ := \sup\{\theta : \varphi_X(\theta) < \infty\}$ and $\theta_- := \inf\{\theta : \varphi_X(\theta) < \infty\}$. One can show that if $\theta_+ > 0$, then $\ln\varphi_X(\cdot)$ is finite and convex on (θ_-,θ_+). For a general function $F : (\theta_-,\theta_+) \to \mathbb{R}$, the function

$$a \mapsto L_F(a) := \sup\{a\theta - F(\theta) : \theta_- < \theta < \theta_+\}$$

is itself convex on (a_-, a_+), with $a_- = \inf\{a : L_F(a) < \infty\}$ and $a_+ = \sup\{a : L_F(a) < \infty\}$. (You can prove these statements as an exercise.) So the next Theorem says that in good cases, the exponential decay rate of $\mathbb{P}(\frac{1}{n}S_n \geqslant a)$ is exactly the Legendre transform of φ_X at a.

(2.23) Cramérs Large Deviation Theorem

Let (X_i) be iid RVs with $\mathbb{E}(X_1) = \mu \in \mathbb{R}$, $S_n = \sum_{i=1}^n X_i$, and assume in addition that

$$(*) \qquad \forall\theta \in \mathbb{R} : \quad \varphi(\theta) = \mathbb{E}(\,\mathrm{e}^{\theta X_1}\,) < \infty.$$

Then for all $a > \mu$,

$$\lim_{n\to\infty}\tfrac{1}{n}\ln\mathbb{P}(\tfrac{1}{n}S_n \geqslant a) = -L_{\ln\varphi}(a) \equiv -\sup\{a\theta - \ln\varphi(\theta) : \theta \in \mathbb{R}\}$$

In the context of (2.17), this means that $\gamma(a) = -L_{\ln \varphi}(a)$.

Proof:

Step 1: It is enough to show the claim when $a = 0$ and $\mu < 0$. Namely, with $Y_i = X_i - a$ we have $\mathbb{E}(Y_1) < 0$, and $\mathbb{P}(S_n \geqslant na) = \mathbb{P}(\sum_{i=1}^n Y_i \geqslant 0)$. On the other hand, $\varphi_{X_1}(\theta) = \mathrm{e}^{a\theta}\varphi_{Y_1}(\theta)$, and therefore $L_{\ln \varphi_{X_1}}(a) = L_{\ln \varphi_{Y_1}}(0)$. This shows the claim of step 1. Note that with $a = 0$ our claim takes a simpler form: we need to prove that

$$(**) \qquad \lim_{n\to\infty} \tfrac{1}{n} \ln \mathbb{P}(S_n \geqslant 0) = \inf\{\ln \varphi(\theta) : \theta \in \mathbb{R}\} = \ln(\inf_{\theta\in\mathbb{R}} \varphi(\theta)).$$

The last equality is because $\varphi \geqslant 0$ and ln is monotone.

Step 2: Assume from now on that $\mathbb{E}(X_1) < 0$ and $a = 0$. By assumption $(*)$, it follows that $\mathbb{E}(|X_1|^p \,\mathrm{e}^{\theta X_1}) < \infty$ for all $p > 0, \theta \in \mathbb{R}$ (exercise!). Therefore, Lebesgues differentiation lemma implies that for all $\theta \in \mathbb{R}$,

$$\varphi'(\theta) = \mathbb{E}(X_1 \,\mathrm{e}^{\theta X_1}) \quad \text{and} \quad \varphi''(\theta) = \mathbb{E}(X_1^2 \,\mathrm{e}^{\theta X_1}) > 0.$$

This shows that φ is strictly convex, and that $\varphi'(0) = \mathbb{E}(X_1) < 0$.

Step 3: In the case that $\mathbb{P}(X_1 \leqslant 0) = 1$, we have $\varphi'(\theta) < 0$ for all θ, and $\lim_{\theta\to\infty} \varphi(\theta) = \mathbb{P}(X_1 = 0)$ (exercise!). Thus in this case,

$$\ln \mathbb{P}(S_n \geqslant 0) = \ln \mathbb{P}(X_i = 0 \,\forall 1 \leqslant i \leqslant n) = \ln \mathbb{P}(X_1 = 0)^n = n \ln \mathbb{P}(X_1 = 0) = n \ln \inf_{\theta} \varphi(\theta),$$

and the claim is shown (both sides above might be $-\infty$).

Step 4: Assume now that $\mathbb{P}(X_1 < 0) > 0$ and $\mathbb{P}(X_1 > 0) > 0$. Then

$$\lim_{\theta\to-\infty} \varphi(\theta) = \lim_{\theta\to\infty} \varphi(\theta) = \infty$$

(exercise!), and thus by strict convexity there exists a unique $\theta_0 \in \mathbb{R}$ with

$$\varphi(\theta_0) = \min\{\varphi(\theta) : \theta \in \mathbb{R}\}, \qquad \text{and} \quad \varphi'(\theta_0) = 0.$$

So in this situation, we therefore just have to show that

$$(***) \qquad \lim_{n\to\infty} \tfrac{1}{n} \ln \mathbb{P}(S_n \geqslant 0) = \ln(\varphi(\theta_0)).$$

Since $\varphi'(0) < 0$, we know in addition that $\theta_0 > 0$. By Chebyshevs inequality, we find that

$$\mathbb{P}(S_n \geqslant 0) \leqslant \mathbb{P}(\mathrm{e}^{\theta_0 S_n} \geqslant 1) \leqslant \mathbb{E}(\mathrm{e}^{\theta_0 S_n}) = \varphi(\theta_0)^n,$$

so we know that $\leqslant$ holds in $(***)$; in fact, we knew this from (2.21) already.

We now need to prove $\geqslant$ in $(***)$ The trick is to change from the random variables X_i where $\mathbb{P}(S_n \geqslant 0)$ vanishes exponentially fast to different random variables $\hat{X}_i$ and $\hat{S}_n = \sum_{i=1}^n \hat{X}_i$ where $\ln \mathbb{P}(\hat{S}_n \geqslant 0)$ remains finite, and to do this in a controlled way so that we can extract the leading order of $\frac{1}{n} \ln \mathbb{P}(S_n \geqslant 0)$ easily from the new expression. The correct way to do this is the **Cramér transform**: Let $\mu = \mathbb{P}_{X_1}$ be the image measure of $\mathbb{P}$ under X. We fix $\hat{\theta} > \theta_0$ and define a probability measure (!) $\hat{\mu}$ by

$$\hat{\mu}(\mathrm{d}x) = \frac{1}{\hat{\rho}} \,\mathrm{e}^{\hat{\theta} x} \,\mu(\mathrm{d}x),$$

where $\hat{\rho} = \varphi(\hat{\theta})$ is the normalization constant. Then let $\hat{X}_1$ be a random variable with distribution $\hat{\mu}$, and observe that its moment generating function $\hat{\varphi}$ is given by

$$\hat{\varphi}(\theta) := \varphi_{\hat{X}_1}(\theta) = \frac{1}{\hat{\rho}} \int_{\mathbb{R}} \mathrm{e}^{\theta x}\ \mathrm{e}^{\hat{\theta} x}\, \mu(\mathrm{d}x) = \frac{1}{\hat{\rho}} \mathbb{E}(\ \mathrm{e}^{(\theta+\hat{\theta})X_1}\ ) = \frac{1}{\hat{\rho}} \varphi_{X_1}(\theta + \hat{\theta}).$$

Therefore, by strict convexity of φ,

$$\varepsilon = \varepsilon(\hat{\theta}) := \mathbb{E}(\hat{X}_1) = \hat{\varphi}'(0) = \frac{1}{\hat{\rho}}\varphi'(\hat{\theta}) > \frac{1}{\hat{\rho}}\varphi'(\theta_0) = 0,$$

and

$$\mathbb{V}(\hat{X}_1) \leqslant \mathbb{E}(\hat{X}_1^2) = \hat{\varphi}''(0) = \frac{1}{\hat{\rho}}\varphi''(\hat{\theta}) = c < \infty,$$

Note that $\varepsilon \to 0$ when $\hat{\theta} \to \theta_0$ by continuity of φ'.

On the other hand,

$$\begin{aligned}\mathbb{P}(S_n \geqslant 0) &= \int_{\{x_1+\ldots+x_n\ \geqslant\ 0\}} \mu(\mathrm{d}x_1) \cdots \mu(\mathrm{d}x_n) = \\ &= \int_{\{x_1+\ldots+x_n\ \geqslant\ 0\}} (\hat{\rho}\, \mathrm{e}^{-\hat{\theta} x_1}\ )\hat{\mu}(\mathrm{d}x_1) \cdots (\hat{\rho}\, \mathrm{e}^{-\hat{\theta} x_n}\ )\hat{\mu}(\mathrm{d}x_n) = \hat{\rho}^n\, \mathbb{E}(\ \mathrm{e}^{-\hat{\theta}\hat{S}_n}\ 1_{\{\hat{S}_n\ \geqslant\ 0\}}),\end{aligned}$$

where $\hat{S}_n = \sum_{i=1}^n \hat{X}_i$. Thus,

$$\begin{aligned}\frac{1}{n}\ln \mathbb{P}(S_n \geqslant 0) &\geqslant \ \ln \hat{\rho} + \frac{1}{n} \ln \mathbb{E}(\ \mathrm{e}^{-\hat{\theta}\hat{S}_n}\ 1_{\{0\ \leqslant\ \hat{S}_n\ \leqslant\ 2n\varepsilon\}}) \geqslant \\ &\geqslant \ \ln \hat{\rho} + \frac{1}{n}\ln\big(\ \mathrm{e}^{-2n\hat{\theta}\varepsilon}\ \mathbb{P}(0 \leqslant \hat{S}_n \leqslant 2n\varepsilon)\big) = \ln\hat{\rho} - 2\hat{\theta}\varepsilon + \frac{1}{n}\ln \mathbb{P}(0 \leqslant \frac{1}{n}\hat{S}_n \leqslant 2\varepsilon)\big).\end{aligned}$$

Since $\mathbb{E}(\hat{X}_1) > 0$ and $\mathbb{V}(X_1) < \infty$, the weak law of large numbers now shows that

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(S_n \geqslant 0) \geqslant \ \ln \hat{\rho} - 2\hat{\theta}\varepsilon = \ln \varphi(\hat{\theta}) - 2\hat{\theta}\varepsilon$$

for all $\hat{\theta} > \theta_0$. Since $\theta \mapsto \varphi(\theta)$ is continuous and $\varepsilon \to 0$ as $\hat{\theta} \to \theta_0$, we can now take the limit $\hat{\theta} \to \theta_0$ and obtain the claim. □

Remark: There are much stronger results on large deviations: it is possible to prove similar results when $\varphi(\theta) < \infty$ only for some θ instead of all, and it is also possible to formulate the statement of Cramérs theorem with the help of so-called *rate functions*; in that form it can be generalized away from just sums of real-valued random variables. We refer to the literature (Klenke, Durrett, Deuschel/Stroock).

Part 4: The Central Limit Theorem, and Convergence of Probability measures

We have treated the CLT in the previous lecture, but we will re-prove it here with a technique that is very important in its own right, namely characteristic functions. We will also use the opportunity to learn many useful things related to the convergence of sequences of probability measures.

(2.24) Motivation

Let (X_n) be iid random variables with $\mathbb{E}(X_1) = 0$, $\mathbb{V}(X_1) = 1$, and let $S_n = \sum_{i=1}^n X_i$. One way to state the weak law of large numbers is to say that

$$\lim_{n\to\infty} \mathbb{P}(\tfrac{1}{n}S_n \in (a,b)) = \begin{cases} 1 & \text{if } 0 \in (a,b) \\ 0 & \text{otherwise} \end{cases}$$

which implies that the sequence of image measures $\mathbb{P}_{\frac{1}{n}S_n}$ converges to the Dirac measure δ_0 at $x = 0$ weakly. Now, the Dirac measure is a very boring probability measure! Why do we not see more interesting things?

Answer: because we killed them all by dividing through n. We should choose a different scaling if we want to have a bit more fun with the limit. Which one should it be? Well,

$$\mathbb{V}(n^{-\alpha}S_n) = n^{-2\alpha}\sum_{i=1}^n \mathbb{V}(X_i) = n^{-2\alpha+1},$$

so the only candidate where the limiting variance is not zero (giving the Dirac measure) or infinity (giving the zero measure or nothing at all, depending on your world view) is $\alpha = 1/2$.

The aim of this part is to prove that the sequence of measures $\mathbb{P}_{n^{-1/2}S_n}$ converges to some non-trivial limiting measure μ, and to identify μ. We start with the 'identify' part: We know that two measures μ and ν on $\mathbb{R}^d$ are equal if $\mu(A) = \nu(A)$ for all rectangles A built from half-open intervals (they form a π-system generating $\mathcal{B}^d$). Alternatively, if we know that

$$(*) \qquad \mathbb{E}_\mu(f) \equiv \int f(\omega)\mu(\mathrm{d}\omega) = \int f(\omega)\nu(\mathrm{d}\omega) = \mathbb{E}_\nu(f)$$

for all $f \in C_b(\mathbb{R}^d)$ (bounded, continuous functions), then we can approximate half-open rectangles from below by such functions and prove (by monotone convergence) that this also implies $\mu = \nu$.

Question: Do we really need $(*)$ for *all* $f \in C_b$?.
Answer: No! A dense subset is enough!

(2.25) Lemma

Let μ, ν be probability measures, and let F be a dense subset of C_b. If $\mathbb{E}_\mu(f) = \mathbb{E}_\nu(f)$ for all $f \in F$, then $\mu = \nu$.

Proof: F is dense, so for $h \in C_b$ and $\varepsilon > 0$ there exists $f \in F$ with $\|f - h\|_\infty < \varepsilon$. Then,

$$|\mathbb{E}_\mu(h) - \mathbb{E}_\nu(h)| \leqslant |\mathbb{E}_\mu(h-f)| + |\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)| + |\mathbb{E}_\nu(f-h)| \leqslant \|h-f\|_\infty + 0 + \|f-h\|_\infty \leqslant 2\varepsilon.$$

Since ε was arbitrary, this shows $\mathbb{E}_\mu(h) = \mathbb{E}_\nu(h)$ for all $h \in C_b$ and so $\mu = \nu$. □

Exercise: Show that (2.25) remains true if we only demand that the linear span of F is dense.

(2.26) Theorem and Definitions

Let μ be a probability measure on $\mathbb{R}$ (or: X a real-valued RV). Then μ (or $\mathbb{P}_X$) is uniquely determined by

a) the sequence

$$\mathbb{N} \ni n \mapsto \int x^n \mu(\mathrm{d}x) \qquad (\text{or } n \mapsto \mathbb{E}(X^n))$$

of its *moments* **if we assume** that $\mu([a,b] = 1)$ (or: $X \in [a,b]$ almost surely) for some $-\infty < a < b < \infty$.

b) its *Laplace transform*

$$\mathbb{R}^+ \ni t \mapsto \int_0^\infty \mathrm{e}^{-tx} \mu(\mathrm{d}x) \qquad (\text{or } t \mapsto \mathbb{E}(\mathrm{e}^{-tX}))$$

if we assume $\mu(\mathbb{R}_0^+) = 1$ (or $X \geqslant 0$ a.s.).

c) its *Fourier transform*

$$\mathbb{R} \ni t \mapsto \int_{-\infty}^\infty \mathrm{e}^{\mathrm{i}tx} \mu(\mathrm{d}x) =: \varphi_\mu(t),$$

or, *characteristic function*

$$\mathbb{R} \ni t \mapsto \varphi_X(t) := \mathbb{E}(\mathrm{e}^{\mathrm{i}tX}).$$

Proof: a) Polynomials are the linear span of the family $F = \{x \mapsto x^n : n \geqslant 0\}$, and they are dense in compact intervals by the Weierstrass theorem.

b) The linear span of $F = \{x \mapsto \mathrm{e}^{-ax} : a \geqslant 0\}$ is dense in $C_b([0,\infty))$ by the Stone-Weierstrass theorem after doing the one-point compactification of $\mathbb{R}^+$ (details: Klenke).

c) will be a consequence of Theorem (2.29) below. □

(2.27) Examples for characteristic functions

a) X with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$, then $\varphi_X(t) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}t} + \mathrm{e}^{-\mathrm{i}t}) = \cos t$.

b) X Poisson(λ)-distributed, then

$$\varphi_X(t) = \mathbb{E}(\mathrm{e}^{\mathrm{i}tX}) = \sum_{k=0}^\infty \mathrm{e}^{\mathrm{i}tk} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t} - 1)}.$$

c) $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\varphi_X(t) = \mathrm{e}^{\mathrm{i}t\mu - \sigma^2 t^2/2}, \quad t \in \mathbb{R}.$$

When you compare this with the Gaussian density $\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, it looks rather similar (at least for $\mu = 0$), but the σ^2 changes place from 'under the fraction line' to 'above the fraction line' in the exponent.

To prove this formula, we calculate

$$\varphi_X(t) = \int_{-\infty}^\infty \mathrm{e}^{\mathrm{i}tx} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2} \mathrm{d}x = \mathrm{e}^{\mathrm{i}\mu t} \int_{-\infty}^\infty \mathrm{e}^{\mathrm{i}(t\sigma)z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \mathrm{d}z,$$

where for the last equality we use the integral substitution $z = (x-\mu)/\sigma$. Since the last expression is equal to $\mathrm{e}^{\mathrm{i}\mu t}\,\varphi_Y(t\sigma)$, with $Y \sim \mathcal{N}(0,1)$, it is enough to show the claimed formula when $\mu = 0, \sigma = 1$. In this case, completing the square in the exponent gives

$$\varphi_Y(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx}\,\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x = \mathrm{e}^{-t^2/2}\,\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-(x-\mathrm{i}t)^2/2}\,\mathrm{d}x$$

The final integral above is equal to one. It is tempting to 'prove' this by the change of variable $z = x - \mathrm{i}t$ which reduces the integral to the standard Gaussian integral $\frac{1}{\sqrt{2\pi}}\int \mathrm{e}^{-z^2/2}\,\mathrm{d}z = 1$. However, change of variable in the complex plane is not that easy and often goes wrong. Here it does work, but the argument must be that we deform the contour of integration (which runs on the line $\{x + \mathrm{i}t : x \in \mathbb{R}\}$ onto the real line, switching back to the original line only for very large x, where the error we make is small. The details are left as an exercise.

d) $X \sim \mathrm{Exp}(\alpha)$, i.e. with density $\rho(x) = \alpha\,\mathrm{e}^{-\alpha x}\,1_{\{x\geqslant 0\}}$. Then

$$\varphi_X(t) = \alpha\int_0^{\infty}\mathrm{e}^{\mathrm{i}tx}\,\mathrm{e}^{-\alpha x}\,\mathrm{d}x = \frac{\alpha}{\alpha - \mathrm{i}t}.$$

(2.28) Properties of the Characteristic Function

Let X be a RV, $\varphi \equiv \varphi_X$ its characteristic function (CF).

a) $\varphi(t) = \mathbb{E}(\cos(tX)) + \mathrm{i}\mathbb{E}(\sin(tX))$, in particular

$$\varphi(0) = 1, \qquad \varphi(-t) = \overline{\varphi(t)}, \qquad |\varphi(t)| \leqslant \mathbb{E}(|\,\mathrm{e}^{\mathrm{i}tX}\,|) = 1.$$

b) $t \mapsto \varphi(t)$ is uniformly continuous on $(-\infty,\infty)$.

Proof: We have

$$\sup_{t\in\mathbb{R}}|\varphi(t+h) - \varphi(t)| = \sup_{t\in\mathbb{R}}\big|\mathbb{E}\big(\underbrace{\mathrm{e}^{\mathrm{i}tX}}_{|.|\,\leqslant\,1}(\,\mathrm{e}^{\mathrm{i}hX}\, - 1)\big)\big| \leqslant \mathbb{E}(|\,\mathrm{e}^{\mathrm{i}hX}\, - 1|).$$

The last expression converges to zero as $h \to 0$ by dominated convergence. □

c) For $a, b \in \mathbb{R}$,

$$\varphi_{aX+b}(t) = \mathbb{E}(\,\mathrm{e}^{\mathrm{i}t(aX+b)}\,) = \mathrm{e}^{\mathrm{i}tb}\,\mathbb{E}(\,\mathrm{e}^{\mathrm{i}(at)X}\,) = \mathrm{e}^{\mathrm{i}tb}\,\varphi_X(at).$$

We have used this in example (2.27) c) in a concrete case.

d) If X and Y are independent, then

$$\varphi_{X+Y}(t) = \mathbb{E}(\,\mathrm{e}^{\mathrm{i}t(X+Y)}\,) = \mathbb{E}(\,\mathrm{e}^{\mathrm{i}tX}\,\mathrm{e}^{\mathrm{i}tY}\,) = \mathbb{E}(\,\mathrm{e}^{\mathrm{i}tX}\,)\mathbb{E}(\,\mathrm{e}^{\mathrm{i}tY}\,) = \varphi_X(t)\varphi_Y(t).$$

This property, which is shared by the Laplace transform and the moment generating function, is very useful!

We can now prove (2.26) c), and more:

(2.29) Theorem: Fourier inversion formula

Let μ be a probability measure on $\mathbb{R}$, and let φ be its characteristic function. Then for all $a < b$,

$$(*) \qquad \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{1}{\mathrm{i}t}\big(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb} \big) \varphi(t)\,\mathrm{d}t = \mu\big((a,b)\big) + \tfrac{1}{2}\big(\mu(\{a\}) + \mu(\{b\})\big),$$

and

$$\mu(\{a\}) = \frac{1}{2T} \lim_{T\to\infty} \int_{-T}^{T} \mathrm{e}^{-\mathrm{i}ta}\,\varphi(t)\,\mathrm{d}t.$$

In particular, if X and Y are RVs, and if $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}$, then $\mathbb{P}_X((a,b]) = \mathbb{P}_Y((a,b])$ for all a, b, and thus X and Y have the same distribution.

Proof: First note that $\frac{1}{\mathrm{i}t}\big(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb} \big) = \int_a^b \mathrm{e}^{-\mathrm{i}ty}\,\mathrm{d}y$, so the absolute value this expression is bounded by $b - a$. We can therefore invoke Fubinis theorem and find

$$I_T := \int_{-T}^{T} \frac{1}{\mathrm{i}t}\big(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb} \big)\varphi(t)\,\mathrm{d}t = \int \mu(\mathrm{d}x) \underbrace{\int_{-T}^{T} \frac{1}{\mathrm{i}t}\big(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb} \big)\,\mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}t}_{=:J_{a,b}(x,T)}.$$

By symmetry,

$$\int_{-T}^{T} \frac{1}{\mathrm{i}t}\,\mathrm{e}^{\mathrm{i}t(x-a)}\,\mathrm{d}t = \int_{-T}^{T} \frac{1}{t}\sin(t(x-a))dt =$$
$$= 2\mathrm{sign}(x-a) \int_0^T \frac{1}{t}\sin(t|x-a|)\,\mathrm{d}t = 2\mathrm{sign}(x-a)\int_0^{T|x-a|} \frac{1}{y}\sin(y)\,\mathrm{d}y.$$

As is often the case, the hardest part of the proof is to show some fun fact about special functions or integrals (aka 'hard analysis'). Here it is the

Claim: For all $R > 0$, $|\int_0^R \frac{1}{y}\sin y\,\mathrm{d}y - \pi/2| \leqslant 2/R$.

Proof of the claim: We have

$$\int_0^R \frac{1}{y}\sin y\,\mathrm{d}y = \int_0^R \mathrm{d}y \int_0^\infty \mathrm{d}z\ \mathrm{e}^{-yz}\,\sin y \overset{\text{Fubini}}{=} \int_0^\infty \mathrm{d}z \int_0^R \mathrm{d}y\ \mathrm{e}^{-yz}\,\sin y = (**).$$

Integration by parts (twice) gives

$$\int_0^R \mathrm{d}y\,\mathrm{e}^{-yz}\,\sin y = -[\,\mathrm{e}^{-yz}\,\cos y]_0^R - \int_0^R \mathrm{d}y z\,\mathrm{e}^{-yz}\,\cos y = -[\,\mathrm{e}^{-yz}\,\cos y]_0^R - [z\,\mathrm{e}^{-yz}\,\sin y]_0^R - \int_0^R \mathrm{d}y z^2\,\mathrm{e}^{-yz}\,\sin y.$$

We rearrange and find

$$\int_0^R \mathrm{d}y\,\mathrm{e}^{-yz}\,\sin y = \frac{1}{1+z^2}(1 - \mathrm{e}^{-zR}\,\cos R - z\,\mathrm{e}^{-zR}\,\sin R).$$

Since

$$\int_0^\infty \frac{1}{1+z^2}\,\mathrm{d}z = [\arctan z]_0^\infty = \pi/2$$

a direct computation now shows the claim.

Continuing with the proof, a simple calculation now shows that

$$\lim_{T\to\infty} J_{a,b}(x,T) = \begin{cases} 2\pi & \text{if } a < x < b \\ \pi & \text{if } x \in \{a,b\} \\ 0 & \text{if } x \notin [a,b]. \end{cases}$$

Now by dominated convergence, $\lim_{T\to\infty} \frac{1}{2\pi} I_T = \mu((a,b)) + \frac{1}{2}\mu(\{a,b\})$. The claim for $\mu(\{a\})$ is proved analogously and left as an exercise. □

We can now easily re-prove the following fact that we already know:

(2.30) Corollary

If $X \sim \mathcal{N}(\mu,\sigma^2)$ and $Y \sim \mathcal{N}(\nu,\kappa^2)$, and if $X \perp\!\!\!\perp Y$, then $X + Y \sim \mathcal{N}(\mu+\nu, \sigma^2+\kappa^2)$.

Proof: exercise. □

(2.31) Central limit theorem with cheat proof

Let (X_n) be iid RVs with $\mathbb{E}(X_1) = 0$, $\mathbb{V}(X_1) = \sigma^2 < \infty$. Then $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ converges in distribution to a $\mathcal{N}(0,\sigma^2)$-distributed random variable Y.

Cheat Proof: Put $S_n = \sum_{i=1}^n X_i$. We do the following computation with the characteristic function of $\frac{1}{\sqrt{n}} S_n$:

$$\begin{aligned} \varphi_{S_n/\sqrt{n}}(t) &= \mathbb{E}(\,\mathrm{e}^{\mathrm{i}\frac{t}{\sqrt{n}}\sum_{i=1}^n X_i}\,) \overset{(2.28d)}{=} \mathbb{E}(\,\mathrm{e}^{\mathrm{i}\frac{t}{\sqrt{n}}X_1}\,)^n \overset{(!!1)}{=} \qquad \text{(Taylor expansion, } \tfrac{t}{\sqrt{n}} \text{ is small)} \\ &\overset{(!!1)}{=} \mathbb{E}\Big(1 + \mathrm{i}\frac{t}{\sqrt{n}}X_1 - \frac{t^2}{2n}X_1^2 + \frac{t^3}{n^{3/2}}R\Big)^n = \Big(1 + 0 - \frac{t^2}{2n}\mathbb{V}(X_1) + \frac{t^3}{n^{3/2}}\mathbb{E}(R)\Big)^n \\ &\overset{(!!2),n\to\infty}{\longrightarrow} \mathrm{e}^{-\frac{\sigma^2}{2}t^2} = \varphi_Y(t). \end{aligned}$$

Since the characteristic functions uniquely characterize the measures, we conclude (!!3) that $\frac{1}{\sqrt{n}}S_n \to Y$ in distribution.

There are several shortcuts (or: cheats) in this proof.

(!!1): The Taylor expansion is not correct since X_1 is a random variable, so even if $t/\sqrt{n}$ is small, $t/\sqrt{n}X_1(\omega)$ might be very large for some ω. Also, we seem to have hidden an X_1^3 in the term R, but we never assumed that $\mathbb{E}(X^3) < \infty$. We will show in Proposition (2.32) below (by Taylor expanding the function $\varphi_{S_n/\sqrt{n}}(t)$ itself instead of $\mathrm{e}^{\mathrm{i}t/\sqrt{n}X_1}$) that the equality with the second expression of the middle line nevertheless holds.

(!!2): This seems trivial. We put $a_n = -t^2/2\mathbb{V}(X_1) + t^3 n^{-1/2}\mathbb{E}(R)$, remember that the logarithm has the Taylor expansion $\ln(1+x) = x - x^2/2 + x^3/3 - \ldots$, and calculate that

$$\ln(1 + \tfrac{a_n}{n})^n = n\ln(1 + \tfrac{a_n}{n}) = n(\tfrac{a_n}{n} + \tfrac{a_n^2}{2n^2} + \ldots) \to \lim_{n\to\infty} a_n = a.$$

But now we realize that a_n is complex valued, and very surprisingly, for complex z the fundamental equality $\ln z^n = n \ln z$ is false in general! The reason is that in $\mathbb{C}$, the logarithm has many branches, and we can not be sure which one we should take. For example, when $z = \mathrm{i}$

and $n = 2$, then $\ln z^2 = \ln(-1)$ which is right on top of the standard branch cut of the ln. We will find a less elegant, but more correct way of proving (!!2) in Theorem (2.33) below and the two Lemmas following it.

(!!3) is where the real trouble starts. We have shown that *equality* of characteristic functions implies equality of distributions. We have *not* shown that *(pointwise) convergence* of CFs implies weak convergence of measures (or: convergence in distribution). To do this is not at all trivial and will keep us busy for the remainder of this chapter (although we will also learn other things along the way).

As promised, we start by solving problem (!!1). We actually do a bit more than that.

(2.32) Proposition

Let X be a real-valued RV.

a) If $\mathbb{E}(|X|^n) < \infty$ for some $n \in \mathbb{N}$, then φ_X is n times differentiable, and

$$\partial_t^k \varphi_X(t) = \mathbb{E}\big((\mathrm{i}X)^k \,\mathrm{e}^{\mathrm{i}tX}\,\big) \quad \forall k \leqslant n.$$

b) If $n \geqslant 2$ in a), then

$$\varphi_X(s) = 1 + \mathrm{i}s\mathbb{E}(X) - \tfrac{s^2}{2}\mathbb{E}(X^2) + \varepsilon(s)s^2$$

with $\lim_{s\to 0} \varepsilon(s) = 0$.

c) Assume that there exists $h > 0$ with

$$(*) \qquad \lim_{n\to\infty} \tfrac{h^n}{n!}\mathbb{E}(|X|^n) = 0.$$

Then φ_X is analytic on $\{z \in \mathbb{C} : |\mathrm{Im} z| < h\}$, and

$$\varphi_X(t+s) = \sum_{k=0}^{\infty} s^k \frac{\mathrm{i}^k}{k!} \mathbb{E}(\,\mathrm{e}^{\mathrm{i}tX}\, X^k) \qquad \forall t \in \mathbb{R}, \forall |s| \leqslant h.$$

Note that the assumption $(*)$ holds e.g. if $\mathbb{E}(\,\mathrm{e}^{|hX|}\,) < \infty$.

Proof: a) Remember that by Taylors theorem, a function f is k times differentiable if and only if there exist numbers $\alpha_1, \ldots \alpha_{k-1}$ so that

$$(**) \qquad f_k(t) := \lim_{h\to 0} \frac{k!}{h^k}\Big(f(t+h) - f(t) - \sum_{j=1}^{k-1} \alpha_j \frac{h^j}{j!}\Big) \quad \text{exists,}$$

and that in this case, $\alpha_j = \partial_t^j f(t)$ and $f_k(t) = \partial_t^k f(t)$. For the function $f = \varphi_X$, the natural guess is that $\alpha_j = \mathbb{E}((\mathrm{i}X)^j \,\mathrm{e}^{\mathrm{i}tX}\,)$. Thus, the expression we need to control as $h \to 0$ is

$$\frac{k!}{h^k}\big(\varphi_X(t+h) - \varphi_X(t) - \sum_{j=1}^{k-1} \mathbb{E}((\mathrm{i}X)^j \,\mathrm{e}^{\mathrm{i}tX}\,)\frac{h^j}{j!}\big) = \mathbb{E}\Big(\underbrace{\frac{k!}{h^k}\,\mathrm{e}^{\mathrm{i}tX}\,(\,\mathrm{e}^{\mathrm{i}hX}\, - \sum_{j=0}^{k-1}\frac{(\mathrm{i}hX)^j}{j!})}_{=:r(k,h,t,X)}\Big).$$

For fixed $x \in \mathbb{R}$, we set $u(x) := \mathrm{e}^{\mathrm{i}hx}$. Further, let $r_1(k,h,t,x) := \operatorname{Re} r(k,h,t,x)$ and $r_2(k,h,t,x) := \operatorname{Im} r(k,h,t,x)$. Then

$$r_1(k,h,t,x) = \operatorname{Re}\left[\frac{k!}{h^k}\,\mathrm{e}^{\mathrm{i}tx}\left(\mathrm{e}^{\mathrm{i}hx} - \sum_{j=0}^{k-1}\frac{1}{j!}x^j(\partial_x^j\,\mathrm{e}^{\mathrm{i}hx}|_{x=0})\right)\right] = \operatorname{Re}\left[\mathrm{e}^{\mathrm{i}tx}\,\frac{k!}{h^k}\frac{1}{k!}x^k\big(\partial_x^k u\big)(\xi_1(x))\right].$$

with $-x < \xi_1(x) < x$. The last equality is because the term in the brackets above is the remainder term when doing the Taylor expansion of the function $x \mapsto u(x)$ up to order $k-1$. Carrying out the differentiation gives that

$$r_1(k,h,t,x) = \operatorname{Re}\left[\mathrm{e}^{\mathrm{i}tx}\,(\mathrm{i}x)^k\,\mathrm{e}^{\mathrm{i}h\xi_1(x)}\right].$$

Analogously, it follows

$$r_2(k,h,t,x) = \operatorname{Im}\left[\mathrm{e}^{\mathrm{i}tx}\,(\mathrm{i}x)^k\,\mathrm{e}^{\mathrm{i}h\xi_2(x)}\right] \text{ with some } -x < \xi_2(x) < x.$$

Therefore $r(k,h,t,X(\omega)) \to \mathrm{e}^{\mathrm{i}tX(\omega)}\,(\mathrm{i}X(\omega))^k$ as $h \to 0$ for all $\omega \in \Omega$, and $|r(k,h,t,X|$ is bounded by $|X|^k \leqslant 1+|X|^n$; the latter is integrable by assumption, and dominated convergence now gives

$$f_k(t) = \lim_{h\to 0}\mathbb{E}(r(h,k,t,X)) = \mathbb{E}((\mathrm{i}X)^k\,\mathrm{e}^{\mathrm{i}tX}).$$

Since $f_k(t) = \partial_t^k\varphi_X(t)$ by $(**)$, a) is shown.

b) We know from $(**)$ (with $t=0$ and $h=s$) and the proof of a) that

$$\frac{2!}{s^2}(\varphi_X(0+s) - 1 - \mathrm{i}s\mathbb{E}(X)) \xrightarrow{s\to 0} \varphi_X''(0) = -\mathbb{E}(X^2).$$

Therefore

$$2\varepsilon(s) := \frac{2!}{s^2}\Big(\varphi_X(s) - \varphi_X(0) - \mathrm{i}s\mathbb{E}(X) + \tfrac{s^2}{2}\mathbb{E}(X^2)\Big) \xrightarrow{s\to 0} 0.$$

Since

$$\varphi_X(s) = 1 + \mathrm{i}s\mathbb{E}(X) - \frac{s^2}{2}\mathbb{E}(X^2) + \varepsilon(s)s^2,$$

b) is shown.

c) By assumption, $\mathbb{E}(|X|^n) < \infty$ for all n, and by a), $|\partial_t^n\varphi_X(t)| \leqslant \mathbb{E}(|X|^n)$ for all n and all t. Therefore by assumption $(*)$, the Taylor series of φ_X around $t \in \mathbb{R}$ has radius of convergence at least h. Thus φ_X can be extended uniquely to a complex analytic function on $\{z \in \mathbb{C} : |\operatorname{Im} z| < h$ with the claimed power series expansion. □

Problem (!!2) in the cheat proof above is solved by the following statement about complex numbers:

(2.33) Proposition

Let $(c_n) \subset \mathbb{C}$ with $\lim_{n\to\infty} c_n = c \in \mathbb{C}$. Then

$$\lim_{n\to\infty}\left(1 + \frac{c_n}{n}\right)^n = \mathrm{e}^c.$$

The proof uses two Lemmata of independent interest. We give them first:

(2.34) Lemma

Let $z_1, \dots, z_n$ and $w_1, \dots, w_n$ be complex numbers with $|z_j| \leqslant \theta$ and $|w_j| \leqslant \theta$ for some $\theta > 0$ and all j. Then

$$|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j| \leqslant \theta^{n-1} \sum_{j=1}^n |z_j - w_j|.$$

Proof: We use induction. For $n = 1$ there is nothing to prove. If the claim holds up to $n-1$, we calculate

$$\Big|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j\Big| \leqslant \Big|z_1 \prod_{j=2}^n z_j - z_1 \prod_{j=2}^n w_j\Big| + \Big|z_1 \prod_{j=2}^n w_j - w_1 \prod_{j=2}^n w_j\Big| \leqslant$$
$$\leqslant \theta \Big|\prod_{j=2}^n z_j - \prod_{j=2}^n w_j\Big| + \theta^{n-1}|z_1 - w_1| \overset{\text{Ind. hyp}}{\leqslant}$$
$$\leqslant \theta\theta^{n-2} \sum_{j=2}^n |z_j - w_j| + \theta^{n-1}|z_1 - w_1| = \theta^{n-1} \sum_{j=1}^n |z_j - w_j|.$$

□

(2.35) Lemma

If $b \in \mathbb{C}$, $|b| \leqslant 1$, then $|\,\mathrm{e}^b - (1+b)| \leqslant |b|^2$.

Proof:

$$|\,\mathrm{e}^b - (1+b)| \leqslant \sum_{k=2}^\infty \frac{|b|^k}{k!} \overset{|b|\leqslant 1}{\leqslant} \frac{|b|^2}{2} \sum_{k=2}^\infty \frac{2}{k!} \leqslant \frac{|b|^2}{2} \sum_{k=0}^\infty 2^{-k} = |b|^2.$$

□

Proof of Proposition 2.33: We use $z_j = (1 + \frac{c_n}{n})$ and $w_j = \mathrm{e}^{c_n/n}$ for all j in Lemma (2.34) and obtain

$$\Big|(1 + \tfrac{c_n}{n})^n - \mathrm{e}^{c_n}\Big| = \Big|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j\Big| \leqslant (\max\{|z_1|, |w_1|\})^{n-1}\, n\, |z_1 - w_1| =: (*).$$

We choose n_0 large enough so that for all $n > n_0$, $|c_n| \leqslant 2|c|$ and $|c_n|/n \leqslant 1$. Then $|z_1| \leqslant 1 + 2|c|/n$, and

$$|w_1| \leqslant |w_1 - z_1| + |z_1| = \big|\,\mathrm{e}^{c_n/n} - (1 + \tfrac{c_n}{n})\big| + \big|1 + \tfrac{c_n}{n}\big| \overset{(2.35)}{\leqslant} \Big(\frac{2|c|}{n}\Big)^2 + 1 + \frac{2|c|}{n}.$$

Therefore,

$$\max\{|z_1|, |w_1|\}^{n-1} \leqslant \Big(1 + \frac{2|c|}{n} + \frac{4|c|^2}{n^2}\Big)^n \overset{n\to\infty}{\longrightarrow} \mathrm{e}^{2|c|},$$

since this time we have $|c| \in \mathbb{R}$ and e.g. the trick with taking the logarithm that failed in the cheat proof will work. Furthermore, we have already just seen that $|w_1 - z_1| \leqslant 2|c|^2/n^2$. Inserting this into $(*)$ shows and taking the limit $n \to \infty$ finishes the proof. □

We now need to solve problem (!!3), which will need much more preparation. We need to investigate weak convergence of probability measures (or: convergence in distribution) more deeply. To be on the safe side, we repeat the definition:

(2.36) Definition

A sequence (μ_n) of probability measures converges *weakly* to a probability measure μ if for all bounded, continuous functions f,

$$\lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu.$$

We write $\mu_n \Rightarrow \mu$. Note that by definition, a sequence of random variables (X_n) converges in distribution to a RV X (we write $X_n \Rightarrow X$) if and only if the sequence of its image measures converges weakly.

(2.37) Proposition

Assume $X_n \Rightarrow X$, and let F_n and F be the distribution function of X_n and X, respectively. Then $F_n(x) \to F(x)$ at all continuity points x of F, i.e. for all x so that F is continuous at x.

Proof: Let

$$g_{x,\varepsilon}(y) = \begin{cases} 1 & \text{if } y \leqslant x, \\ 1 - \frac{y-x}{\varepsilon} & \text{if } x < y < x+\varepsilon, \\ 0 & \text{if } y > x+\varepsilon. \end{cases}$$

Then $g_{x,\varepsilon} \in C_b(\mathbb{R})$, and

$$\mathbb{P}(X_n \leqslant x) = \mathbb{E}(1_{(-\infty,x]}(X_n) g_{x,\varepsilon}(X_n)) \leqslant \mathbb{E}(g_{x,\varepsilon}(X_n)),$$

and therefore

$$\limsup_{n\to\infty} F_n(x) \leqslant \limsup_{n\to\infty} \mathbb{E}(g_{x,\varepsilon}(X_n)) = \mathbb{E}(g_{x,\varepsilon}(X)) \leqslant \mathbb{P}(X \leqslant x+\varepsilon) = F(x+\varepsilon).$$

Since distribution functions are always continuous from the right, we can take $\varepsilon \to 0$ and find $\limsup_{n\to\infty} F_n(x) \leqslant F(x)$ for all x. Similarly,

$$\liminf_{n\to\infty} F_n(x) \geqslant \liminf_{n\to\infty} \mathbb{E}(g_{x-\varepsilon,\varepsilon}(X_n)) = \mathbb{E}(g_{x-\varepsilon,\varepsilon}(X)) \geqslant \mathbb{P}(X \leqslant x-\varepsilon) = F(x-\varepsilon).$$

If x is a continuity point of F, we can again take $\varepsilon \to 0$ and find $\liminf_{n\to\infty} F_n(x) \geqslant F(x)$. □

The next theorem is somewhat surprising: in (1.28) we have seen that convergence in distribution is in some sense the weakest form of convergence, it does not imply any of the others. However, if the X_n are real-valued, and if we are prepared to abandon the probability space that we are given for another one, we can make weakly convergent RVs even almost surely convergent, which is quite a strong form of convergence. In detail:

(2.38) Theorem

Let (X_n), X be real-valued RVs with $X_n \Rightarrow X$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and RVs (Y_n), Y on Ω with the properties that

a) $Y_n \sim X_n$ for all n, and $Y \sim X$ ($\sim$ means they have the same distribution).

b) $Y_n \to Y$ almost surely.

Proof: Let F_n be the distribution function of X_n for all $n \leqslant \infty$, with $X_\infty := X$. We put

$$\Omega := (0,1), \quad \mathbb{P} := \lambda_{(0,1)}, \quad Y_n(\omega) := \sup\{x \in \mathbb{R} : F_n(x) < \omega\}.$$

We will show that

a) For all n, Y_n has distribution function F_n, so $Y_n \sim X_n$.

b) A monotone function has at most countably many jumps.

c) For all ω such that Y_∞ is continuous at ω, we have $\lim_{n\to\infty} Y_n(\omega) = Y_\infty(\omega)$.

Once we have proved a) - c), the claim holds since then outside the countable (therefore measure zero) set where Y_∞ has jumps, the random variables Y_n converge.

Proof of a):

$$\begin{aligned} Y_n(\omega) \leqslant z \;&\Leftrightarrow\; \sup\{x \in \mathbb{R} : F_n(x) < \omega\} \leqslant z \overset{F_n \text{ monotone}}{\Leftrightarrow} F_n(z + \tfrac{1}{m}) \geqslant \omega \;\forall m \in \mathbb{N} \\ &\Leftrightarrow \omega \in (0, F_n(z + \tfrac{1}{m})] \;\forall m \in \mathbb{N} \;\Leftrightarrow\; \omega \in (0, F_n(z)]. \end{aligned}$$

Therefore,

$$\mathbb{P}(Y_n \leqslant z) = \lambda(0,1)((0, F_n(z)]) = F_n(z) = \mathbb{P}(X_n \leqslant z).$$

Proof of b): Let $f : \mathbb{R} \to \mathbb{R}$ be monotone increasing. For $x \in \mathbb{R}$, let

$$a_x := \sup\{f(z) : z < x\}, \quad b_x := \inf\{f(z) : z > x\}.$$

Then $a_x \leqslant b_x \leqslant a_{x+\varepsilon}$ for all $\varepsilon > 0$ and all x, and therefore $(a_x, b_x) \cap (a_z, b_z) = \emptyset$ whenever $x \neq z$. The set of discontinuities of f is given by

$$A := \{x \in \mathbb{R} : a_x < b_x\} = \{x \in \mathbb{R} : (a_x, b_x) \neq \emptyset\}.$$

If A is empty, the function f is continuous on $\mathbb{R}$ and the claim is true. Otherwise, pick $q_x \in (a_x, b_x) \cap \mathbb{Q}$ for each $x \in A$. Since the intervals (a_x, b_x) are disjoint for different $x \in A$, this gives a one-to-one map from A to a subset of $\mathbb{Q}$. Thus A is countable.

Proof of c): Let ω be such that Y_∞ is continuous at ω. Then
(i): $F_\infty(y) < \omega$ if $y < Y_\infty(\omega)$ by definition of Y_∞,
(ii): $F_\infty(y) > \omega$ if $y > Y_\infty(\omega)$. To see this, assume that $F_\infty(y) = \omega$, then $Y_\infty(\omega + \varepsilon) = \sup\{x \in \mathbb{R} : F_\infty(x) < \omega + \varepsilon\} \geqslant y$ for all $\varepsilon > 0$. This would imply that Y_∞ is discontinuous at ω, which contradicts our assumption.

Now pick $y < Y_\infty(\omega)$ such that F_∞ is continuous at y. Such an y exists since F_∞ has only countably many jumps. By Proposition (2.37), $\lim_{n\to\infty} F_n(y) = F_\infty(y)$, and by (i),

$$\exists n_0 \in \mathbb{N} : \forall n > n_0 : F_n(y) < \omega.$$

This means that $\liminf_{n\to\infty} Y_n(\omega) \geqslant y$ for all y such that F_∞ is continuous at y. Since there are only countably many y where this is not the case, we conclude $\liminf_{n\to\infty} Y_n(\omega) \geqslant Y_\infty(\omega)$.

Finally, pick $y > Y_\infty(\omega)$ such that F_∞ is continuous at y. Then $F_n(y) \to F_\infty(y)$ as $n \to \infty$, and by (ii) and the same argument as above, $\limsup_{n\to\infty} Y_n(\omega) \leqslant y$. This holds for all $y > Y_\infty(\omega)$ where $F_\infty(y)$ is continuous, and since there are only countably many y where this is not the case, we conclude $\limsup_{n\to\infty} Y_n(\omega) \leqslant Y_\infty(\omega)$. This shows c). □

(2.39) Corollary: Fatou's Lemma for weak convergence

Let $g : \mathbb{R} \to \mathbb{R}$ be continuous and assume $g(x) \geqslant 0$ for all x. Let X_n, X be real-valued random variables. If $X_n \Rightarrow X$, then

$$\liminf_{n\to\infty} \mathbb{E}(g(X_n)) \geqslant \mathbb{E}(g(X)).$$

Proof: Exercise. □.

Now we can prove the converse of Proposition (2.37).

(2.40) Proposition

Let (X_n), $n \leqslant \infty$ be real RVs, F_n their distribution functions. Assume that $\lim_{n\to\infty} F_n(x) = F_\infty(x)$ whenever x is a continuity point of F_∞. Then $X_n \overset{n\to\infty}{\Rightarrow} X$.

Proof: Choose (Y_n) with $Y_n \sim X_n$ and $Y_n \to Y_\infty$ almost surely as given by (2.38). Then for each $g \in C_b(\mathbb{R})$, it follows that $g(Y_n) \to g(Y_\infty)$ a.s. by continuity of g, and thus by dominated convergence, $\mathbb{E}(g(X_n)) = \mathbb{E}(g(Y_n)) \to \mathbb{E}(g(Y_\infty)) = \mathbb{E}(g(X_\infty))$. □

(2.41) Continuous Mapping Theorem

Let $g : \mathbb{R} \to \mathbb{R}$ be measurable, and define

$$D_g := \{x \in \mathbb{R} : g \text{ is not continuous at } x\}.$$

(i): If for random variables X_n, X we have $X_n \overset{n\to\infty}{\Rightarrow} X$, and if $\mathbb{P}(X \in D_g) = 0$, then we have $g(X_n) \overset{n\to\infty}{\Rightarrow} g(X)$.

(ii): If in addition to the conditions of (i) g is bounded, then also $\mathbb{E}(g(X_n)) \overset{n\to\infty}{\to} \mathbb{E}(g(X))$.

Proof: (i): We need to show that for each $f \in C_b(\mathbb{R})$,

$$(*) \qquad \lim_{n\to\infty} \mathbb{E}(f(g(X_n))) = \mathbb{E}(f(g(X)).$$

As in the previous proof, choose Y_n with $Y_n \sim X_n$, Y with $Y \sim X$, and $Y_n \to Y$ almost surely. Since f is continuous, we have $D_{f\circ g} \subset D_g$, and so $\mathbb{P}(Y \in D_{f\circ g}) \leqslant \mathbb{P}(Y \in D_g) = \mathbb{P}(X \in D_g) = 0$. Consequently, the set

$$\Omega_0 := \{\omega \in \Omega : Y_n(\omega) \not\to Y(\omega)\} \cup \{\omega \in \Omega : f \circ g \text{ is discontinuous at } Y(\omega)\}$$

has measure zero. For $\omega \notin \Omega_0$, we have $\lim_{n\to\infty} f(g(Y_n(\omega))) \to f(g(Y(\omega)))$, and thus by dominated convergence (and the fact that $Y_n \sim X_n$, $Y \sim X$) we conclude $(*)$.

(ii): For bounded g, we are allowed to take $f(x) = x$ in (i) since $f \circ g$ is still bounded. This gives the conclusion. □

We now need another basic theorem about weak convergence. It collects equivalent criteria that can be used to detect weak convergence of measures and has a strange name. We formulate it for random variables, but you can easily translate it to statements for the image measures.

(2.42) Portmanteau Theorem

Let X, X_n be RVs. The following statements are equivalent:

(i): $X_n \Rightarrow X$.

(ii): $\liminf \mathbb{P}(X_n \in G) \geqslant \mathbb{P}(X \in G)$ for all open subsets G of $\mathbb{R}$.

(iii): $\limsup_{n\to\infty} \mathbb{P}(X_n \in K) \leqslant \mathbb{P}(X \in K)$ for all closed subsets K of $\mathbb{R}$.

(iv): For all $A \in \mathcal{B}(\mathbb{R})$ with $\mathbb{P}(X \in \partial A) = 0$, we have $\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$.
(∂A is the boundary of A, i.e. in each little ball around a point of ∂A we find both points in and outside of A).

Remark: The inequalities in (ii) and (iii) can be strict. To see this, take $\mathbb{P}(X_n = 1/n) = 1$, $G = (0,1)$ and $K = [-1, 0]$.

Proof:

(i) ⇒ (ii): Choose $Y_n \sim X_n$, $Y \sim X$, and $Y_n \to Y$ almost surely. If $Y_n(\omega)$ converges, G is open and $Y(\omega) \in G$, then $Y_n(\omega) \in G$ for all large enough n. Thus $\liminf_{n\to\infty} 1_G(Y_n) \geqslant 1_G(Y)$ almost surely, and so

$$\liminf \mathbb{P}(Y_n \in G) = \liminf \mathbb{E}(1_G(Y_n)) \overset{\text{Fatou}}{\geqslant} \mathbb{E}(\liminf 1_G(Y_n)) \geqslant \mathbb{E}(1_G(Y)) = \mathbb{P}(Y \in G).$$

By $X_n \sim Y_n$, $X \sim Y$ this then holds for the X_n, X as well.

(ii) ⇔ (iii): We have $\{G : G \subset \mathbb{R} : G \text{ open}\} = \{K^c : K \subset \mathbb{R}, K \text{ closed}\}$, and thus

$$\liminf \mathbb{P}(X_n \in G) \geqslant \mathbb{P}(X \in G) \Leftrightarrow \limsup \underbrace{(1 - \mathbb{P}(X_n \in G))}_{\mathbb{P}(X_n \in G^c)} \leqslant \underbrace{1 - \mathbb{P}(X \in G)}_{\mathbb{P}(X \in G^c)}$$

shows the equivalence.

(ii) and *(iii) ⇒ (iv)*: For $A \in \mathcal{B}(\mathbb{R})$, let A° be the open interior of A and $\bar{A}$ be the closure of A. Then $\partial A = \bar{A} \setminus A^\circ$. By assumption, then $\mathbb{P}(X \in A^\circ) = \mathbb{P}(X \in A) = \mathbb{P}(X \in \bar{A})$. Now (ii) implies

$$\liminf \mathbb{P}(X_n \in A) \geqslant \liminf \mathbb{P}(X_n \in A^\circ) \geqslant \mathbb{P}(X \in A^\circ) = \mathbb{P}(X \in A),$$

and (iii) implies

$$\limsup \mathbb{P}(X_n \in A) \leqslant \limsup \mathbb{P}(X_n \in \bar{A}) \leqslant \mathbb{P}(X \in \bar{A}) = \mathbb{P}(X \in A).$$

Together this gives (iv).

(iv) ⇒ (i): Let F be the distribution function of X and let x be a continuity point of F. Then

$$0 = \mathbb{P}(X \in \{x\}) = \mathbb{P}(X \in \partial(-\infty, x]),$$

and so (iv) implies

$$F_n(x) = \mathbb{P}(X_n \leqslant x) \to \mathbb{P}(X \leqslant x) = F(x).$$

By (2.40), this implies weak convergence. □

(2.43) Reformulation of (2.37), (2.40), and some comments

Let X_n be RVs, F_n their distribution functions. The following statements a) and b) are equivalent:

a) There exists a RV X with $X_n \Rightarrow X$.

b) There exists a nondecreasing, right-continuous function F such that
  (i): $F_n(x) \to F(x)$ at all continuity points of F.
  (ii): F is the distribution function of some RV X.

Question: Is b(i) already enough to imply a)?

Answer: No! As a counterexample, let G be a continuous distribution function, and let $a, b, c \in \mathbb{R}^+$ with $a + b + c = 1$. Let

$$F_n(x) = a1_{[n,\infty)}(x) + b1_{[-n,\infty)}(x) + cG(x),$$

then each F_n is a distribution function, and $F_n(x) \overset{n\to\infty}{\longrightarrow} b + cG(x) =: F_\infty(x)$. But F_∞ is not a distribution function since $\lim_{x\to-\infty} F_\infty(x) = b \neq 0$ and $\lim_{x\to\infty} F_\infty(x) = b + c = 1 - a \neq 1$. In terms of random variables, this means that if X_n has distribution function F_n, and if there would be a RV X with $X_n \Rightarrow X$, this would mean that for all $N \in \mathbb{R}^+$,

$$\mathbb{P}(X \in (-N, N)) \overset{(2.42)(ii)}{\leqslant} \liminf \mathbb{P}(X_n \in (-N, N)) = c(G(N) - G(-N)).$$

By continuity from below ((1.3) c), this would imply that $\mathbb{P}(X \in \mathbb{R}) \leqslant c < 1$, which is impossible.

What has happened? Mass has escaped to infinity! Explicitly, note that the image measure of X_n has a point mass of size a at $n \in \mathbb{R}$, and another point mass of size b at $-n \in \mathbb{R}$, and these wander off to infinity as $n \to \infty$. So while the weak limit of the image measures is still a measure, it is no longer a probability measure as its mass is smaller than 1.

In the next statements we will see that this is the only bad thing that can happen: if we exclude it, b(i) is already enough. Even better, we will see that b(i) holds **along a subsequence** for *any* sequence of distribution functions. We do this first.

(2.44) Hellys Selection Theorem

Let (F_n) be a sequence of distribution functions. Then there exists a subsequence $(F_{n_k})_{k\in\mathbb{N}}$ and a non-decreasing, right-continuous function F such that $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ for all continuity points of F.

Proof: Let (q_j) be an enumeration of $\mathbb{Q}$. Since $(F_n(q_1))_{n\in\mathbb{N}} \subset [0, 1]$ and the latter set is compact, there is an increasing $\mathbb{N}$-valued sequence $(n_1(k))_{k\in\mathbb{N}}$ so that $\lim_{k\to\infty} F_{n_1(k)}(q_1)$ exists. Then, since also $(F_{n_1(k)}(q_2))_{k\in\mathbb{N}} \subset [0, 1]$, there is a subsequence $(n_2(k))$ of the sequence $(n_1(k))_{k\in\mathbb{N}}$ so that also $\lim_{k\to\infty} F_{n_2(k)}(q_2)$ exists; of course, the first limit still exists along this subsequence. You recognize Cantors diagonal argument: by induction, we now find sequence of sequences $(n_j(k))_{k\in\mathbb{N}}$ so that each $(n_j(k))$ is a subsequence of $(n_{j-1}(k))$ and such that $\lim_{k\to\infty} F_{n_j(k)}(q_i)$ exists for all $i \leqslant j$. Choosing the sequence $(n_j(j))_{j\in\mathbb{N}}$ ensures that $G(q) := \lim F_{n_j(j)}(q)$ exists for all $q \in \mathbb{Q}$. We define

$$F(x) := \inf\{G(q) : q \in \mathbb{Q} : q > x\}.$$

Then F is right continuous since

$$\lim_{x_m \searrow x} F(x_m) = \inf\{G(q) : q \in \mathbb{Q}, \exists m \text{ with } q > x_m\} = \inf\{G(q) : q \in \mathbb{Q} : q > x\} = F(x).$$

F is also increasing since $\{G(q) : q \in \mathbb{Q}, q > x\} \supset \{G(q) : q \in \mathbb{Q}, q > y\}$ if $x < y$, and so the infimum of the larger set is smaller or equal than the one of the smaller set.
Finally $F_{n_j(j)}(x) \to F(x)$ at all continuity points x of F: to see this, choose x so that F is continuous at x. For given $\varepsilon > 0$ pick $r_1, r_2, s \in \mathbb{Q}$ with $r_1 < r_2 < x < s$ and

$$F(x) - \varepsilon < F(r_1) \leqslant F(r_2) \leqslant F(x) \leqslant F(s) < F(x) + \varepsilon.$$

This is possible since F is continuous at x. Now we have $F_{n_j(j)}(r_2) \to G(r_2) \geqslant F(r_1)$, and $F_{n_j(j)}(s) \to G(s) \leqslant F(s)$. So for all sufficiently large j, we find

$$F(x) - \varepsilon < F_{n_j(j)}(r_2) \leqslant F_{n_j(j)}(x) \leqslant F_{n_j(j)}(s) < F(x) + \varepsilon$$

It follows that for each $\varepsilon > 0$,

$$F(x) - \varepsilon \leqslant \liminf_{j\to\infty} F_{n_j(j)}(x) \leqslant \limsup_{j\to\infty} F_{n_j(j)}(x) \leqslant F(x) + \varepsilon.$$

Taking $\varepsilon \to 0$ shows the convergence and hence the result. □

Now we treat the issue of escape of mass. First we give a name to the situation where escape of mass is excluded.

(2.45) Definition

a) A sequence (μ_n) of probability measures on $\mathbb{R}$ is *tight* if

$$\forall\varepsilon > 0 : \exists K \subset \mathbb{R}, K \text{ compact, such that } \mu_n(K) \geqslant 1 - \varepsilon \quad \forall n \in \mathbb{N}.$$

b) A sequence (X_n) of real-valued RVs is *tight* if the sequence $(\mathbb{P}_{X_n})$ of image measures is tight; explicitly, if

$$\forall\varepsilon > 0 : \exists K \subset \mathbb{R}, K \text{ compact, such that } \mathbb{P}(X_n \in K) \geqslant 1 - \varepsilon \quad \forall n \in \mathbb{N}.$$

c) A sequence of distribution functions (F_n) is *tight* if

$$\forall\varepsilon > 0 : \exists M > 0 : \liminf_{n\to\infty} F_n(M) - F_n(-M) > 1 - \varepsilon.$$

It is an easy exercise to show that (X_n) is tight iff the sequence of distribution functions (F_n) is tight.

(2.46) Theorem

Let (X_n) be a sequence of RVs, (F_n) the sequence of their distribution functions. The following statements are equivalent:

a) (X_n) is tight, equivalently (F_n) is tight.

b) For every convergent subsequence (F_{n_k}) of (F_n), there exists a RV X so that $\lim_{k\to\infty} F_{n_k}$ is the distribution function of X.

c) If (X_{n_k}) is a subsequence of (X_n) so that $\lim_{k\to\infty} \mathbb{P}(X_{n_k} \in (-\infty, y])$ exists for all $y \in \mathbb{R}$, then there exists a RV X with $X_{n_k} \overset{k\to\infty}{\Rightarrow} X$.

Proof: The equivalence *b)* ⇔ *c)* is by definition of distribution functions.

a) ⇒ *b)*: Assume that (X_n) is tight and that (F_{n_k}) is a convergent subsequence of (F_n). By using (2.44), we can conclude that there is a further subsequence $(F_{n_{k_l}})$ of (F_{n_k}) and a non-decreasing and right-continuous function F so that $\lim_{l\to\infty} F_{n_{k_l}}(x) = F(x)$ at all continuity points of F. But since (F_{n_k}) itself was convergent, also $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ at all continuity points of F. By tightness, we can for $\varepsilon > 0$ fix M so that $\mathbb{P}(|X_n| > M) < \varepsilon$ for all n. Let $r < -M$ and $s > M$ be continuity points of F. They exist since F only has countably many discontinuities. Then

$$F(s) - F(r) = \lim_{k\to\infty} F_{n_k}(s) - F_{n_k}(r) \geqslant \limsup_{k\to\infty}(F_{n_k}(M) - F_{n_k}(-M)) \geqslant 1 - \varepsilon.$$

Here, the first inequality holds by monotonicity of the F_{n_k}, and the second by our choice of M. Again since there are only countably many discontinuities, we conclude that $\liminf_{x\to\infty} F(x) - F(-x) = 1$, so F is the distribution function of some RV X.

b) ⇒ *a)*: Assume that (F_n) is not tight. Then there is $\varepsilon > 0$ and a subsequence (F_{n_k}) with

$$F_{n_k}(k) - F_{n_k}(-k) \leqslant 1 - \varepsilon \qquad \forall k \in \mathbb{N}.$$

By (2.44), there exists a non-decreasing, right-continuous function F so that a further subsequence of (F_{n_k}) converges to F at the continuity points of F. Then as above, we see that $F(s) - F(r) \leqslant 1 - \varepsilon$ for all continuity points of F, so F cannot be a distribution function. □

By combining the above theorems, we can formulate concisely:

(2.47) Theorem

a) A sequence (μ_n) of probability measures on $\mathbb{R}$ has a subsequence that converges weakly to some probability measure if and only if (μ_n) is tight.

b) A sequence of random variables (X_n) has a subsequence that converges in distribution to some random variable iff (X_n) is tight.

Now we can finally come back to the task of proving (!!3). Before we do this, let me recommend the great book 'Convergence of Probability measures' by P. Billingsley for much more interesting stuff about weak convergence of measures. In particular, the restriction to real valued RVs is completely unnecessary if we do things correctly.

The result that solves problem (!!3) is

(2.48) Lévy's Continuity Theorem

Let (μ_n) be a sequence of probability measures on $\mathbb{R}$, and (φ_n) their characteristic functions.

(i): Assume that $\mu_n \Rightarrow \mu_\infty$ for some probability measure μ_∞, and let φ_∞ be the characteristic function of μ_∞. Then $\varphi_n(t) \to \varphi_\infty(t)$ for all $t \in \mathbb{R}$.

(ii): Assume that there exists a function $\varphi_\infty : \mathbb{R} \to \mathbb{C}$, with the properties that $\varphi_n(t) \to \varphi_\infty(t)$ for all $t \in \mathbb{R}$, and that φ_∞ is continuous at $t = 0$. Then φ_∞ is the characteristic function of some probability measure μ_∞, and $\mu_n \Rightarrow \mu_\infty$.

Proof: (i): Since $\varphi_n(t) = \int \mathrm{e}^{\mathrm{i}tx}\, \mu_n(\mathrm{d}x)$ and the function $x \mapsto \mathrm{e}^{\mathrm{i}tx}$ is continuous, this holds simply by the definition of weak convergence applied to real and imaginary parts.

(ii): First we show that the sequence (μ_n) is tight under the stated assumptions. Fix $\varepsilon > 0$; we have to find $K \in \mathbb{R}$ so that $\liminf_{n\to\infty} \mu_n([-K,K]) \geqslant 1-\varepsilon$. For all $K > 0$ and all functions h that are positive on $[-K,K]^c$, Chebyshevs inequality gives

$$\mu_n([-K,K]^c) \leqslant \frac{1}{\inf\{h(x) : |x| > K\}} \int_{[-K,K]^c} h(x)\mu_n(\mathrm{d}x).$$

The function that makes the connection to the characteristic function is $h(x) = 1 - \frac{\sin(x/K)}{x/K}$. Indeed, it is not hard to see that for all $x \geqslant 1$, $\sin(x)/x \leqslant \alpha := \sin(1)/1$, and so $h(x) \geqslant 1-\alpha$ for all $x > K$. So,

$$\mu_n([-K,K]^c) \leqslant \frac{1}{1-\alpha} \int_{[-K,K]^c} h(x)\mu_n(\mathrm{d}x).$$

On the other hand,

$$h(x) = 1 - \tfrac{K}{x}\sin(x/K) = 1 - \int_0^1 \cos(tx/K)\,\mathrm{d}t = \int_0^1 (1-\cos(tx/K)\,\mathrm{d}t,$$

and so by Fubini's Theorem (everything is nonnegative!), we find that

$$\int_{[-K,K]^c} h(x)\mu_n(\mathrm{d}x) = \int_0^1 \mathrm{d}t \int_{\mathbb{R}} (1-\cos(tx/K))\mu_n(\mathrm{d}x) = \int_0^1 \big(1 - \mathrm{Re}\,\varphi_n(t/K)\big)\,\mathrm{d}t.$$

We assumed convergence of characteristic functions, and hence dominated convergence gives

$$\limsup_{n\to\infty} \mu_n([-K,K]^c) \leqslant \frac{1}{1-\alpha}\int_0^1 \big(1 - \mathrm{Re}\,\varphi_\infty(t/K)\big)\,\mathrm{d}t.$$

We also assumed continuity of φ_∞ at $t = 0$, which means that we can bring $\inf\{\varphi_\infty(t/K) : 0 \leqslant t \leqslant 1\}$ as close to $\varphi_\infty(0) = 1$ as we want by taking K large. This shows tightness.

By tightness we know that (μ_n) converges to some μ_∞ along a subsequence. We still have to show convergence without a subsequence. This follows by a standard argument from analysis: pick any subsequence (μ_{n_k}) of (μ_n). Then by tightness, there exists a further subsequence $(\mu_{n_{k(j)}})_j$ that converges to some probability measure ν. The characteristic function of ν is given by φ_∞: to see this, note that the CF of ν is the limit of the sequence of characteristic functions $(\varphi_{n_{k(j)}})_j$ by part (i), and the limit of this sequence is φ_∞ by the assumed convergence of CFs. Since the CF uniquely determines the measure, we conclude that $\nu = \mu_\infty$. In other words every subsequence of (μ_n) has a further subsequence that converges to μ_∞. The standard argument from analysis is that in such a case, the sequence itself must already converge. To be explicit, assume that (μ_n) does not converge to μ_∞. Then we can pick $\delta > 0$, a subsequence $(\mu_{n(i)})_i$ and a continuous and bounded function f so that for all i, $|\int f(x)\mu_{n(i)}(\mathrm{d}x) - \int f(x)\mu_\infty(\mathrm{d}x)| > \delta$. Applying the argument above to that particular subsequence gives a contradiction. Hence $\mu_n \Rightarrow \mu_\infty$. □

Now we can finally really prove the CLT:

(2.49) Central Limit Theorem for iid random variables

Let (X_i) be iid real RVs with $\mathbb{E}(X_i) = \mu \in \mathbb{R}$ and $\mathbb{V}(X_i) = \sigma^2 \in (0, \infty)$ for all i. Then

$$\frac{1}{\sqrt{\sigma^2 n}} \sum_{i=1}^{n} (X_i - \mu) \Rightarrow Y \qquad \text{with } Y \sim \mathcal{N}(0,1).$$

Proof: By considering $X_i' = X_i - \mu$, we only need to treat the case $\mu = 0$. By (2.32 b),

$$\varphi_{X_1}(t) = 1 - \frac{\sigma^2 t^2}{2} + \varepsilon(t) t^2$$

with $\lim_{t \to 0} \varepsilon(t) = 0$. Setting $S_n = \sum_{i=1}^n X_i$ and using independence gives

$$\varphi_{S_n/\sqrt{\sigma^2 n}}(t) = \mathbb{E}\big(\mathrm{e}^{\mathrm{i}\frac{t}{\sqrt{\sigma^2 n}} \sum_{j=1}^n X_j}\big) = \big(\varphi_{X_1}(\tfrac{t}{\sqrt{n\sigma^2}})\big)^n = \big(1 - \tfrac{t^2}{2n} + \tfrac{t^2}{n\sigma^2}\varepsilon(\tfrac{t}{\sqrt{n\sigma^2}})\big)^n \overset{(2.33)}{\longrightarrow} \mathrm{e}^{-t^2/2}\,.$$

By (2.48), this shows the claim. □

(2.50) Remarks

a) Consider the situation of (2.49) with $\mu = 0$. Define

$$\bar{S}_n(\omega) := \frac{1}{\sqrt{n\sigma^2}} \sum_{i=1}^{n} X_i(\omega)$$

This random variable converges in distribution to $\mathcal{N}(0,1)$ as we have just seen, but it does not converge almost surely. Even more drastically, the set of ω where $\bar{S}_n(\omega)$ does converge has measure zero. To see this, note the following:

(i): Since the normalization $1/\sqrt{n}$ kills all finite sums of X_i, we have for each $M > 0$:

$$\{\limsup_{n\to\infty} \bar{S}_n \geqslant M\} = \bigcap_{k \in \mathbb{N}} \{ \limsup_{n\to\infty, n>k} \tfrac{1}{\sqrt{n\sigma^2}} \sum_{i=k+1}^{n} X_i \geqslant M\} \in \mathcal{T}, \text{ the tail } \sigma\text{-algebra.}$$

The same is true for $\{\liminf_{n\to\infty} \bar{S}_n \leqslant -M\}$, and so the probability that $\bar{S}_n$ oscillates between arbitrarily large and arbitrarily small values infinitely often is either zero or one by Kolmogorov's $0-1$-law.

(ii): We will now show that zero is not an option in (i). For each $M > 0$, we have

$$\{\limsup_{n\to\infty} \bar{S}_n \geqslant M\} = \bigcap_{n\in\mathbb{N}} \bigcup_{m \geqslant n} \underbrace{\{\bar{S}_m \geqslant M\}}_{=:A_m} = \limsup_{n\to\infty} A_n,$$

and by Fatou's Lemma applied to $1 - 1_{A_n}$, we find

$$\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n \geqslant M) = \mathbb{P}(\limsup_{n\to\infty} A_n) \geqslant \limsup_{n\to\infty} \mathbb{P}(A_n) = \frac{1}{\sqrt{2\pi}} \int_M^\infty \mathrm{e}^{-x^2/2}\,\mathrm{d}x > 0.$$

By (i), this only leaves the possibility that $\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n \geqslant M) = 1$, and the same holds for $\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n \leqslant -M)$. Thus $\bar{S}_n$ converges with probability zero.

On the other hand, (2.49) and (2.38) tell us that there exists a sequence of RVs (Y_n) with $Y_n \sim \bar{S}_n$ that converges almost surely. But the 'natural' choice $\bar{S}_n$ behaves very badly for pointwise convergence!

b) Pairwise independence (i.e. $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$) is not enough to prove the CLT: as an

example, consider a sequence (ξ_i) of iid Bernoulli RVs, i.e. $\mathbb{P}(\xi_i = \pm 1) = 1/2$. We are looking for random variables (X_i) with the properties that
(i): $X_i \perp\!\!\!\perp X_j$ if $i \neq j$, and $X_i \sim \xi_1$ for all i.
(ii): $\sum_{i=1}^{2^{n-1}} X_i = \xi_1(1+\xi_2)(1+\xi_3)\cdots(1+\xi_n)$ for all n.

Assume we have found such X_i. Then we write $\bar{S}_m = \frac{1}{\sqrt{\frac{m}{2}}} \sum_{i=1}^{\frac{m}{2}} X_i$. Property (i) means that $\mathbb{E}(X_i) = 0$ and $\mathbb{V}(X_i) = 1$, and so $\bar{S}_m$ is the right candidate for a CLT.

By property (ii), however, $\mathbb{P}(\bar{S}_{2^n} = 0) = 1 - 2^{-n+1}$. This means that for bounded continuous g, $\mathbb{E}(g(\bar{S}_{2^n})) \to \mathbb{E}(g(0))$, so along that particular subsequence the sequence $\bar{S}_m$ converges in distribution to a Dirac measure at zero. This excludes the validity of the CLT.

It remains to find X_i with the properties (i) and (ii). We can take inspiration from (ii): the sum on the left hand side has 2^{n-1} terms, and there are also exactly 2^{n-1} terms if we multiply out the right hand side; so those terms are strong candidates for the X_i . We just have to order them correctly: we start with just $X_1 = \xi_1$; we have used the 1's in all the remaining brackets after ξ_1 here. The second is the additional term that we get when we use the ξ_2 instead of the 1 in the second bracket. I.e. $X_2 = \xi_1\xi_2$. We now have two terms. The next two terms are obtained by multiplying each of them by ξ_3, i.e. by using the ξ_3 in the third bracket instead of the 1 on everything that we have obtained so far. The result is $X_3 = \xi_1\xi_3$ and $X_4 = \xi_1\xi_2\xi_3$. The next four terms are obtained by multiplying each of the four terms that we already have by ξ_4, and so on. You should now check as an exercise that these X_i have the property (i), which finishes the example.

We have just seen that we cannot replace independence by pairwise independence. What we can do, however, is to relax the condition of identical distribution.

(2.51) Lindeberg-Feller CLT

For each $n \in \mathbb{N}$, let $(X_{n,m})_{m \leqslant n}$ be a family of independent RVs. Assume that for all m, n $\mathbb{E}(X_{n,m}) =: \mu_{n,m}$ and $\mathbb{V}(X_{n,m}) =: \sigma_{n,m}^2$ exist. Define

$$s_n^2 := \sum_{m=1}^{n} \sigma_{n,m}^2,$$

and assume that for all $\varepsilon > 0$,

$$(*) \qquad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{m=1}^{n} \mathbb{E}\big((X_{n,m} - \mu_{n,m})^2 1_{\{|X_{n,m}-\mu_{n,m}| \geqslant \varepsilon s_n\}} = 0.$$

(the so-called Lindeberg condition). Then

$$\frac{1}{s_n^2} \sum_{m=1}^{n} (X_{n,m} - \mu_{n,m}) \overset{n\to\infty}{\Rightarrow} Y \qquad \text{with } Y \sim \mathcal{N}(0,1).$$

Proof: See the books of Durett or Klenke. The Lindeberg condition means that no single random variable can dominate the sum: note that without the indicator function, the expression in $(*)$ would be equal to 1. Also note that in the identically distributed case, $s_n = 1$. So indeed, the Lindeberg condition means that no single $\sigma^2_{n,m}$ can contribute a finite amount to s_n as $n \to \infty$.

3. Conditional Expectation

You probably recall the conditional probability $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$, defined if $\mathbb{P}(B) > 0$. The following extension of this concept is one of the most important definitions of modern probability.

(3.1) Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and X be a real-valued RV with $\mathbb{E}(|X|) < \infty$. Let $\mathcal{G}$ be a σ-algebra over Ω with $\mathcal{G} \subset \mathcal{F}$ (i.e. $\mathcal{G}$ contains fewer sets). Any random variable $Y : \Omega \to \mathbb{R}$ is called **conditional expectation of X given $\mathcal{G}$** if it has the following two properties:

(i): Y is $\mathcal{G}$-measurable

(ii): For all $G \in \mathcal{G}$, the equality $\mathbb{E}(X1_G) = \mathbb{E}(Y1_G)$ holds.

If Y is a conditional expectation of X given $\mathcal{G}$, it is customary to write $Y = \mathbb{E}(X \,|\, \mathcal{G})$.

(3.2) Remarks

a) This definition is very convenient to work with, but not that easy to understand intuitively. We will see examples below that should help you understand it.

b) We will later see that the conditional expectation is unique almost surely, i.e. if both Y and $\tilde{Y}$ fulfill (i) and (ii) above, then $Y(\omega) = \tilde{Y}(\omega)$ outside a set of measure zero. This justifies to just write $\mathbb{E}(X \,|\, \mathcal{G})$.

c) A formula that you should memorize is

$$\mathbb{E}(Z\mathbb{E}(X \,|\, \mathcal{G})) = \mathbb{E}(ZX) \quad \text{for all bounded, } \mathcal{G}\text{- measurable RVs } Z.$$

The proof is by one of the most frequently used arguments of probability theory: for $Z = 1_G$, $G \in \mathcal{G}$, this holds by definition. For finite linear combinations $Z = \sum_{i=1}^n \alpha_i 1_{G_i}$ with $G_i \in \mathcal{G}$, it then holds by linearity of $\mathbb{E}$. For positive X and Z, it holds by monotone convergence. For general X and Z, it holds by decomposition into positive and negative parts and linearity again. If you are not completely familiar with the procedure, you should now do the details as an exercise.

We will now give examples to illustrate Definition (3.1).

(3.3) Examples

a) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $A_i \in \mathcal{F}$, $i = 1, \ldots, n$, be a partition of Ω, i.e. $A_i \cap A_j = \emptyset$ if $i \neq j$, and $\bigcup_{i=1}^n A_i = \Omega$. Assume $\mathbb{P}(A_i) \neq 0$ for all i. Let $\mathcal{G} = \sigma(\{A_i : 1 \leqslant i \leqslant n\})$.

Let X be any RV on Ω with $\mathbb{E}(|X|) < \infty$. Then

$$Y : \Omega \to \mathbb{R}, \quad \omega \mapsto \sum_{i=1}^{n} 1_{A_i}(\omega) \frac{\mathbb{E}(X 1_{A_i})}{\mathbb{P}(A_i)}$$

is a conditional expectation of X given $\mathcal{G}$ (exercise!). $Y(\omega)$ is the average of X on the set A_i containing ω, and is therefore the best approximation to (or best guess of) X when we only know in which A_i we find ω, but not the exact value of ω. This leads to the important intuition of conditional expectation as a best approximation under incomplete information.

b) In example a), let $\Omega = (0,1]$ with Borel-σ-algebra and Lebesgue measure, $k \in \mathbb{N}$, $A_i = ((i-1)2^{-k}, i2^{-k}]$, with $1 \leqslant i \leqslant 2^k$. For an integrable RV X, we then have

$$\mathbb{E}(X|\mathcal{G})(\omega) = \sum_{i=1}^{2^k} 1_{\{(i-1)2^{-k} < \omega \leqslant i2^{-k}\}} \frac{1}{2^{-k}} \int_{(i-1)2^{-k}}^{i2^{-k}} X(s) \mathrm{d}s.$$

So, this is a step function, similar to the one used in the Riemann-Approximation to the integral, but 'better' since instead of the left, right or trapeze rule value, the 'true' average of X is used on each Riemann interval.

c) In example a), let $X = 1_B$ for $B \in \mathcal{F}$. Then $\mathbb{E}(X|\mathcal{G})$ encodes all classical conditional probabilities $\mathbb{P}(B|A_i)$ at once: if $\omega \in A_i$, then $\mathbb{E}(1_B|\mathcal{G}) = \mathbb{P}(B|A_i)$.

d) The real strength of Definition (3.1) is that it is not restricted to σ-Algebras $\mathcal{G}$ that are generated by finitely many sets (and thus are finite). Let $\Omega = [-L, L]^2$ with Borel-σ-algebra and normalized Lebesgue measure. Let

$$\mathcal{G} = \sigma(\{\{(u,v) : -L \leqslant v \leqslant L, a \leqslant u \leqslant b\} : -L \leqslant a \leqslant b \leqslant L\})$$

be the σ-algebra generated by the 'vertical cylinders'. Then for any integrable RV X,

$$Y(u,v) := \frac{1}{2L} \int_{-L}^{L} X(u,w) \, \mathrm{d}w$$

is a conditional expectation of X given $\mathcal{G}$ (exercise!). Again, this is the best guess of the true value $X(u,v)$ if we know that the first coordinate is u but have no information about the second coordinate. From this example, we also see that $\mathbb{E}(X|\mathcal{G})$ is not unique: setting $Y(u,v) = 0$ for all u in a set of Lebesgue-measure zero will also give a conditional expectation (exercise).

e) The construction of d) works for more fancy choices of $\mathcal{G}$ too: think of polar coordinates; or let Z be any random variable and let $\mathcal{G} = \sigma(Z)$, see (1.12). The last situation is very important and we will come back to it later.

(3.4) Lemma

Let X be an integrable, $\mathcal{F}$-measurable RV, $\mathcal{G} \subset \mathcal{F}$ and $Y = \mathbb{E}(X|\mathcal{G})$ a conditional expectation of X given $\mathcal{G}$. Then Y is integrable.

Proof: Let $A^+ = \{Y \geqslant 0\}$ and $A^- = \{Y < 0\}$. By property (i) of Def. (3.1), $A^+, A^- \in \mathcal{G}$. Thus, by property (ii),

$$\mathbb{E}(|Y|) = \mathbb{E}(1_{A^+} Y) + \mathbb{E}(1_{A^-} Y) = \mathbb{E}(1_{A^+} X) + \mathbb{E}(1_{A^-} X) \leqslant 2\mathbb{E}(|X|) < \infty$$

□

We now prove existence and uniqueness of the conditional expectation. As usual, uniqueness is easier, so we start with it.

(3.5) Proposition

In the situation of Definition (3.1), assume that Y and $\tilde{Y}$ both have properties (i) and (ii). Then $Y = \tilde{Y}$ $\mathbb{P}$-almost surely.

Proof: For all $G \in \mathcal{G}$,

$$\mathbb{E}((Y - \tilde{Y})1_G) = \mathbb{E}(Y 1_G) - \mathbb{E}(\tilde{Y} 1_G) \overset{(ii)}{=} \mathbb{E}(X 1_G) - \mathbb{E}(X 1_G) = 0.$$

Since $Y - \tilde{Y} \in m\mathcal{G}$, we thus have

$$\mathbb{E}((Y - \tilde{Y})1_{\{Y \geqslant \tilde{Y}\}}) = 0.$$

Since $(Y - \tilde{Y})1_{\{Y \geqslant \tilde{Y}\}}$ is integrable (by (3.4)) and nonnegative, monotonicity of the expected value (see (1.14) e) gives $(Y - \tilde{Y})1_{\{Y \geqslant \tilde{Y}\}} = 0$ almost surely. The same argument shows that $(Y - \tilde{Y})1_{\{Y < \tilde{Y}\}} = 0$ almost surely, finishing the proof. □

(3.6) About existence of conditional expectation

The existence of conditional expectation can be proved using the Radon-Nikodym theorem. But we will take another, more geometric approach, which gives additional insight into the concept. We will show that the conditional expectation $\mathbb{E}(X|\mathcal{G})$ of some square integrable RV is the **orthogonal projection** of the vector $X \in L^2(\mathrm{d}\mathbb{P})$ **onto the closed subspace of** $L^2(\mathrm{d}\mathbb{P})$ **consisting of all $\mathcal{G}$-measurable, square integrable RVs.** As usual, we work with equivalence classes of $\mathbb{P}$-almost surely equal functions. We introduce some notions from functional analysis. If you are unfamiliar with them, you should study them. Alternatively, you can take the existence of conditional expectation as a black box.

(3.7) Definitions and Facts from the theory of L^2 spaces

We define

$$\mathscr{L}^2_{\mathcal{F}}(\mathbb{P}) = \{X \in m\mathcal{F} : \mathbb{E}(|X|^2) < \infty\}.$$

You should check that this is a vector space. Note the restriction to $\mathcal{F}$-measurable functions. We say that $X \in \mathscr{L}^2_{\mathcal{F}}$ is equivalent to $Y \in \mathscr{L}^2_{\mathcal{F}}$ and write $X \sim Y$ if $X = Y$ $\mathbb{P}$-almost surely. Then $L^2_{\mathcal{F}}(\mathbb{P}) = L^2$ is the vector space whose points (vectors) are equivalence classes in $\mathscr{L}^2_{\mathcal{F}}(\mathbb{P})$. We define, for $X, Y \in L^2_{\mathcal{F}}(P)$ the inner product

$$(X, Y) := \mathbb{E}(\overline{X}_0 Y_0),$$

where $\overline{X}$ is complex conjugation, and X_0, Y_0 are representatives of the equivalence classes X and Y. You should check that the definition is independent of the choice of representative, that $X \mapsto (X, Y)$ is an antilinear map from L^2 to $\mathbb{C}$ and that $Y \mapsto (X, Y)$ is a linear such map.

Also, $\|X\| := (X,X)^{1/2}$ is a norm on L^2, and L^2 is complete under this norm (i.e. Cauchy sequences converge). This (by definition) makes L^2 a **Hilbert space**. As usual, we will often denote equivalence classes and their representatives with the same symbol.

(3.8) Orthogonal projection in Hilbert spaces

Let V be a Hilbert space, U a closed subspace of V, $x \in V$. Then for $y \in U$, the following statements are equivalent:
(i): $\|y - x\| \leqslant \|z - x\|$ for all $z \in U$.
(ii): $(y - x, w) = 0$ for all $w \in U$.

If (i) or (and) (ii) hold, then y is called **orthogonal projection** of x onto U.

Proof: Let $y \in U$. Since $U = \{y + w : w \in U\}$, we have

$$\begin{aligned} \forall z \in U : \|y-x\|^2 \leqslant \|z-x\|^2 &\Leftrightarrow \forall w \in U : \|y-x\|^2 \leqslant \|y+w-x\|^2 \\ &\Leftrightarrow \forall w \in U : \|y-x\|^2 \leqslant \|y-x\|^2 + \|w\|^2 + 2\mathrm{Re}\,(y-x,w) \\ &\Leftrightarrow \forall w \in U : 0 \leqslant \|w\|^2 + 2\mathrm{Re}\,(y-x,w). \qquad (*) \end{aligned}$$

So (ii) implies (i). On the other hand, assume that (ii) is false. Then we can find $w \in U$ with $\mathrm{Re}\,(y-x,w) < 0$ (why?). Now,

$$\|\varepsilon w\|^2 + 2\mathrm{Re}\,(y-x,\varepsilon w) = \varepsilon^2\|w\|^2 + 2\varepsilon \mathrm{Re}\,(y-x,w) < 0$$

for small enough ε. So, $(*)$ is false for this εw, and thus (i) does not hold. □

(3.9) Proposition

Let V be a Hilbert space, U a closed subspace of V. Then for each $x \in V$, there exists a unique orthogonal projection of x onto U.

Proof: If $x \in U$, then x is the unique vector y that fulfills (i) of (3.8) and is therefore the projection. For $x \notin U$, put $d := \inf\{\|w - x\| : w \in U\}$. By the definition for an infimum, there exists a sequence $(w_n) \in U$ such that $\lim_{n\to\infty} \|w_n - x\| = d$. We calculate, for $m, n \in \mathbb{N}$,

$$\|w_n - x\|^2 + \|w_m - x\|^2 = 2\|\tfrac{1}{2}(w_n + w_m) - x\|^2 + 2\|\tfrac{1}{2}(w_n - w_m)\|^2.$$

Since $\frac{1}{2}(w_n + w_m) \in U$, the first term on the right hand side above is $\geqslant 2d^2$ for all m, n. We conclude

$$\lim_{N\to\infty} \sup_{m,n>N} \tfrac{1}{2}\|w_n - w_m\|^2 \leqslant \lim_{N\to\infty} \Big(\sup_{n>N} \|w_n - x\|^2 + \sup_{m>N} \|w_m - x\|^2 - 2d^2 \Big) = d^2 + d^2 - 2d^2 = 0.$$

So, (w_n) is a Cauchy sequence. Since U is closed, its limit is in U, and it fulfils (3.8) (i). □

(3.10) Theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Then $L^2_{\mathcal{G}}(\mathbb{P})$ is a closed subspace of $L^2_{\mathcal{F}}(\mathbb{P})$. For any RV X with $\mathbb{E}(|X|^2) < \infty$, every representative of the orthogonal projection of X onto $L^2_{\mathcal{G}}$ is a conditional expectation of X given $\mathcal{G}$.

Proof: The claim that $L^2_{\mathcal{G}}$ is a closed subspace is left as an easy exercise. Let $X \in L^2_{\mathcal{F}}$ and let Y be the orthogonal projection onto $L^2_{\mathcal{G}}$. Since Y is an equivalence class of $\mathcal{G}$-measurable functions, every representative fulfils (i) of Definition (3.1). For (ii), let $G \in \mathcal{G}$. Then

$$\mathbb{E}(1_G Y - 1_G X) = (Y - X, 1_G) = 0$$

by property (3.8) (ii). □

Before we extend this to L^1, is is convenient to state some simple properties of conditional expectation.

(3.11) Proposition: integral-related properties of conditional expectation

Let X, Y be integrable RVs, and assume that all conditional expectations below exist.

a) (Linearity): We have, for almost all ω,

$$\mathbb{E}(\alpha X + \beta Y|\mathcal{G})(\omega) = \alpha \mathbb{E}(X|\mathcal{G})(\omega) + \beta \mathbb{E}(Y|\mathcal{G})(\omega).$$

b) (Monotonicity): If $X \geqslant 0$ a.s., then also $\mathbb{E}(X|\mathcal{G}) \geqslant 0$ a.s..

Proof: a) is just easy definition chasing. For b), let $Y = \mathbb{E}(X|\mathcal{G})$ and $A = \{\omega : Y(\omega) \leqslant 0\}$. Then $A \in \mathcal{G}$, and we have

$$0 \overset{X \geqslant 0}{\leqslant} \mathbb{E}(X 1_A) \overset{(3.1)(ii)}{=} \mathbb{E}(Y 1_A).$$

Since $Y 1_A \leqslant 0$, we conclude $Y 1_A = 0$ a.s., thus $Y \geqslant 0$ a.s. □

(3.12) Theorem

Let X be an integrable RV and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Then $\mathbb{E}(X|\mathcal{G})$ exists.

Proof: Assume first that $X \geqslant 0$, and let $X_n = X \wedge n$. Then X_n is bounded and thus in L^2, and Theorem (3.10) gives the existence of $Y_n = \mathbb{E}(X_n|\mathcal{G})$. By (3.11) b), the sequence (Y_n) is monotone. Let $Y := \lim_{n\to\infty} Y_n$ a.s. We have that $Y \in m\mathcal{G}$, and monotone convergence gives for all $G \in \mathcal{G}$

$$\mathbb{E}(Y 1_G) = \lim_{n\to\infty} \mathbb{E}(Y_n 1_G) \overset{(3.1)(ii)}{=} \lim_{n\to\infty} \mathbb{E}(X_n 1_G) = \mathbb{E}(X 1_G).$$

For general X, decompose into positive and negative part, use the above argument and recompose by linearity. □

In the following statements, we always assume that all occurring RVs are integrable, and that all curly letters denote σ-algebras.

(3.13) Proposition: basic measurability properties

a) Expected value is conserved: $\mathbb{E}\big(\mathbb{E}(X|\mathcal{G})\big) = \mathbb{E}(X)$

b) $\mathcal{G}$-measurable functions do not change: if $X \in m\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G})(\cdot) = X(\cdot)$ a.s.

c) The 'Tower property': If $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then

$$\mathbb{E}\big(\mathbb{E}(X|\mathcal{G})\big|\mathcal{H}\big) = \mathbb{E}(X|\mathcal{H}) \qquad \text{a.s.}$$

Proof: For a) use (3.1) (ii) with $G = \Omega$.
For b) note that X already fulfils (3.1) (i) and (ii).
For c), we have to check the definition for X replaced by $\mathbb{E}(X|\mathcal{G})$ and Y replaced by $\mathbb{E}(X|\mathcal{H})$. (3.1) (i) is clear since $\mathbb{E}(X|\mathcal{H}) \in m\mathcal{H}$. For (ii), let $H \in \mathcal{H} \subset \mathcal{G}$. Then,

$$\mathbb{E}(1_H \mathbb{E}(X|\mathcal{G})) \overset{(3.1)\underline{(ii)} \text{ for } \mathcal{G}}{=} \mathbb{E}(1_H X) \overset{(3.1)\underline{(ii)} \text{ for } \mathcal{H}}{=} \mathbb{E}(1_H \mathbb{E}(X|\mathcal{H})).$$

□

(3.14) Proposition: limits and inequalities

a) Monotone convergence: if $0 \leqslant X_n \nearrow_{n\to\infty} X$ a.s., then also $0 \leqslant \mathbb{E}(X_n|\mathcal{G}) \nearrow_{n\to\infty} \mathbb{E}(X|\mathcal{G})$ a.s.

b) Conditional Fatou: if $0 \leqslant X_n$ for all n, then

$$\mathbb{E}(\liminf_{n\to\infty} X_n|\mathcal{G})(\cdot) \leqslant \liminf_{n\to\infty} \mathbb{E}(X_n|\mathcal{G})(\cdot) \quad \text{a.s.}$$

c) Dominated convergence: Assume that $|X_n(\cdot)| \leqslant Z(\cdot)$ for all n for some integrable Z a.s., and that $X_n(\cdot) \to X(\cdot)$ a.s. Then

$$\lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{G})(\cdot) = \mathbb{E}(X|\mathcal{G})(\cdot) \quad \text{a.s.}$$

d) Conditional Jensen: Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and assume $\mathbb{E}(|\varphi(X)|) < \infty$. Then

$$\mathbb{E}(\varphi(X)|\mathcal{G})(\cdot) \geqslant \varphi\big(\mathbb{E}(X|\mathcal{G})(\cdot)\big) \quad \text{a.s}$$

e) L^p-contractivity: If $X \in L^p$, $1 \leqslant p \leqslant \infty$, then $\mathbb{E}(X|\mathcal{G}) \in L^p$, and

$$\|\mathbb{E}(X|\mathcal{G})\|_{L^p}^p \equiv \mathbb{E}\big(|\mathbb{E}(X|\mathcal{G})|^p\big) \leqslant \mathbb{E}\big(|X|^p\big) \equiv \|X\|_{L^p}^p.$$

Proof: a) As in the proof of Theorem (3.12), we see that $Y := \lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{G})$ a.s. satisfies the definition of a conditional expectation of X given $\mathcal{G}$.

b) Let $Y_n := \inf_{k \geqslant n} X_k$. By monotonicity of conditional expectation, $\mathbb{E}(Y_n|\mathcal{G}) \leqslant \mathbb{E}(X_k|\mathcal{G})$ a.s. for all $k \geqslant n$, and so

$$\mathbb{E}(\inf_{k \geqslant n} X_k|\mathcal{G}) = \mathbb{E}(Y_n|\mathcal{G}) \leqslant \inf_{k \geqslant n} \mathbb{E}(X_k|\mathcal{G}).$$

As $n \to \infty$, the expression on the left hand side converges to $\mathbb{E}(\liminf_{n\to\infty} X_n|\mathcal{G})$ by a). The expression on the right hand side converges to $\liminf_{n\to\infty} \mathbb{E}(X_n|\mathcal{G})$.

c) By assumption, $Z + X_n \geqslant 0$ a.s. By b), a.s.,

$$\mathbb{E}(Z + X|\mathcal{G}) = \mathbb{E}(Z + \liminf_{n\to\infty} X_n|\mathcal{G}) \leqslant \liminf_{n\to\infty} \mathbb{E}(Z + X_n|\mathcal{G}),$$

and by linearity of conditional expectation this shows $\mathbb{E}(X|\mathcal{G}) \leqslant \liminf \mathbb{E}(X_n|\mathcal{G})$ a.s. As $Z - X_n \geqslant 0$ for all n a.s. is also true, the same argument shows $-\mathbb{E}(X|\mathcal{G}) \leqslant -\liminf \mathbb{E}(X_n|\mathcal{G})$, thus $\mathbb{E}(X|\mathcal{G}) \geqslant \limsup \mathbb{E}(X_n|\mathcal{G})$ a.s.

d) Convexity implies that the left derivative $\partial_-\varphi(x)$ exists for all x, and that the graph of φ is always above the tangent at $\varphi(x)$ with slope $\partial_-\varphi(x)$. In symbols:

$$\forall x \in \mathbb{R}, \forall y \in \mathbb{R} : \varphi(x) + \partial_-\varphi(x)(y - x) \leqslant \varphi(y).$$

Since φ is convex, it is continuous, and so

$$\forall y \in \mathbb{R} : \varphi(y) = \sup\{\varphi(q) + (y-q)\partial_-\varphi(q) : q \in \mathbb{Q}\};$$

By choosing an enumeration (c_n) of $\mathbb{Q}$ and putting $a_n := \partial_-\varphi(c_n)$, $b_n := \varphi(c_n) - c_n\partial_-\varphi(c_n)$, we find that

$$(*) \forall y \in \mathbb{R} : \varphi(y) = \sup\{a_n y + b_n : n \in \mathbb{N}\}.$$

By using this inequality for $y = X(\omega), \omega \in \Omega$, we find that

$$\forall \omega \in \Omega, \forall n \in \mathbb{N} : \varphi(X(\omega)) \geqslant a_n X(\omega) + b_n.$$

For fixed $n \in \mathbb{N}$, we can now take conditional expectation and use monotonicity to find that for each n, there exists a set $\Omega_n \subset \Omega$ with $\mathbb{P}(\Omega_n) = 1$ and

$$\mathbb{E}(\varphi(X)|\mathcal{G})(\omega) \geqslant a_n \mathbb{E}(X|\mathcal{G})(\omega) + b_n \qquad \forall \omega \in \Omega_n$$

This shows that for $\omega \in \bigcap_{n\in\mathbb{N}} \Omega_n$,

$$\mathbb{E}(\varphi(X)|\mathcal{G})(\omega) \geqslant \sup\{a_n \mathbb{E}(X|\mathcal{G})(\omega) + b_n : n \in \mathbb{N}\} \overset{(*)}{=} \varphi(\mathbb{E}(X|\mathcal{G})(\omega)).$$

Since $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \Omega_n) = 1$, the claim is proved.
(Exercise: look at the (much simpler) proof of the ordinary Jensen inequality, e.g. from last semester, and find out why it cannot be easily adapted to work for conditional expectation.)

e) From d) with $\varphi(x) = |x|^p$, $1 \leqslant p < \infty$, we get

$$|\mathbb{E}(X|\mathcal{G})|^p \leqslant \mathbb{E}\big(|X|^p \big| \mathcal{G}\big) \qquad \text{a.s..}$$

The claim follows by taking expectation and using (3.13 a). The claim for $p = \infty$ follows from monotonicity:

$$\pm\mathbb{E}(X|\mathcal{G}) \leqslant \mathbb{E}(|X|\,|\mathcal{G}) \quad \text{a.s.} \qquad \Longrightarrow \qquad |\mathbb{E}(X|\mathcal{G})| \leqslant \mathbb{E}(|X|\,|\mathcal{G}) \quad \text{a.s.}$$

The claim now follows by taking the essential supremum over ω; essential supremum means that we may leave out a set of measure zero when taking the supremum. □

(3.15) Proposition: Advanced measurability properties

a) 'Measurable factors can be pulled out of the conditional expectation':
If Z is a RV with $ZX \in L^1$ and $Z \in m\mathcal{G}$, then

$$\mathbb{E}(XZ|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G}) \quad \text{a.s.}$$

b) 'Independent information is irrelevant':
If $\mathcal{H} \subset \mathcal{F}$ and $\mathcal{H} \perp\!\!\!\perp \sigma(\mathcal{G}, \sigma(X))$, then

$$\mathbb{E}(X|\sigma(\mathcal{G}, \mathcal{H})) = \mathbb{E}(X|\mathcal{G}).$$

Note that the condition on $\mathcal{H}$ is stronger than the condition that $\mathcal{H}$ is independent from $\mathcal{G}$ and from $\sigma(X)$.

c) A special case of b): if $\mathcal{H} \perp\!\!\!\perp \sigma(X)$, then $\omega \mapsto \mathbb{E}(X|\mathcal{H})$ is a.s. constant and equal to $\mathbb{E}(X)$.

Proof: a) We check the definition: for (i), we confirm that $\omega \mapsto Z(\omega)\mathbb{E}(X|\mathcal{G})(\omega)$ is $\mathcal{G}$-measurable.

For (ii), we start with $X \geqslant 0$ and $Z = 1_A$ for some $A \in \mathcal{G}$. Then for $B \in \mathcal{G}$, $A \cap B \in \mathcal{G}$, and we have

$$(*) \qquad \mathbb{E}(1_B Z \mathbb{E}(X|\mathcal{G})) = \mathbb{E}(1_{A\cap B}\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(1_{A\cap B}X) = \mathbb{E}(1_B Z X),$$

which is (ii). By using the standard method of linarity and monotone convergence (we still have $X \geqslant 0$), we find that $(*)$ holds for all nonnegative Z fulfilling our assumptions. Now we decompose X and Z into positive and negative parts and get four terms on both sides of $(*)$ so that for each of them both random variables have a definite sign. Linearity now concludes the proof.

b) Exercise. □

(3.16) Mnemonic rule: the smaller σ-algebra always wins

If $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then

$$\mathbb{E}\Big(\mathbb{E}(X|\mathcal{G})\Big|\mathcal{H}\Big) = \mathbb{E}\Big(\mathbb{E}(X|\mathcal{H})\Big|\mathcal{G}\Big) = \mathbb{E}(X|\mathcal{H}).$$

Proof: Exercise. □

In many applications, the σ-algebra $\mathcal{G}$ is generated by a random variable. This case is so important that it gets its own notation.

(3.17) Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\Omega', \mathcal{F}')$ a measurable space, and $Z : \Omega \to \Omega'$ be a RV. Recall that $\sigma(Z) := \{Z^{-1}(A) : A \in \mathcal{F}'\}$. For each integrable real RV X, the real RV

$$\mathbb{E}(X|Z) : \Omega \to \mathbb{R}, \qquad \omega \mapsto \mathbb{E}(X|Z)(\omega) := \mathbb{E}(X|\sigma(Z))(\omega)$$

is called the conditional expectation of X given Z.

(3.18) Example

$(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$, $\Omega' = \mathbb{N}$, $\mathcal{F}' = \mathcal{P}(\Omega')$. For a RV $Z : \Omega \to \Omega'$ define $A_i := Z^{-1}(\{i\})$. Then the (A_i) form a partition of Ω. Let $X : \Omega \to \mathbb{R}$ be any integrable RV.

Assume that Ω describes an experiment, and ω is the true state of the physical system that we want to measure. Our measurement apparatus only allows us to evaluate $Z(\omega)$, and we are interested in predicting the value of $X(\omega)$. Once we have measured $Z(\omega)$ and the result is $j \in \mathbb{N}$, then the best guess for $X(\omega)$ is $\mathbb{E}(X|Z)(\omega)$ for some $\omega \in A_j$. The function $\omega \mapsto \mathbb{E}(X|Z)(\omega)$ is constant on all the A_j, and equal to $\mathbb{E}(X|A_j)$. There seems to be a problem for those A_j with $\mathbb{P}(A_j) = 0$, but in this case $Z(\omega) = j$ will never be measured, and we can assign any value to $\mathbb{E}(X|Z)(\omega)$ for those ω. The RV $\mathbb{E}(X|Z)$ assigns the 'best guess for $X(\omega)$ given $Z(\omega)$' to each value of ω by the procedure

(i): determine $A_{Z(\omega)}$, i.e. the set of all $\tilde{\omega} \in \Omega$ which are indistinguishable from ω when only the information provided by Z is available,

(ii): compute the weighted average of X over $A_{Z(\omega)}$.

(3.19) Example

Let X, Z be real RVs so that the distribution of the RV $(X, Z) : \Omega \to \mathbb{R}^2$ has a density $\rho_{X,Z} : \mathbb{R}^2 \to \mathbb{R}_0^+$. Recall that this means in particular that

$$\mathbb{P}(X \in A, Z \in B) = \int_A \mathrm{d}x \int_B \mathrm{d}z \, \rho_{X,Z}(x, z) \qquad \forall A, B \in \mathcal{B}(\mathbb{R}).$$

You should check the validity of the following statements:
(i): The maps $x \mapsto \rho_{X,Z}(x, z)$ (for fixed z) and $z \mapsto \rho_{X,Z}(x, z)$ (for fixed x) are usually not densities of a probability measure.
(ii): The map $x \mapsto \rho_X(x) := \int \rho_{X,Z}(x, z) \, \mathrm{d}z$ is a density for $\mathbb{P} \circ X^{-1}$, and the map $z \mapsto \rho_Z(z) := \int \rho_{X,Z}(x, z) \, \mathrm{d}x$ is a density for $\mathbb{P} \circ Z^{-1}$.
(iii): For each z, the map

$$x \mapsto f_{X|Z}(x; z) := \frac{1}{\rho_Z(z)} \rho_{X,Z}(x, z) 1_{\{\rho_Z(z) > 0\}}$$

is the density of the distribution of a RV. It is called the *conditional density of X given that $Z = z$.*
(iv): For each $\mathbb{P}_X$-integrable function h, the map

$$\omega \mapsto \int h(x) \rho_{X|Z}(x, Z(\omega)) \, \mathrm{d}x$$

is a version of $\mathbb{E}(h(X)|Z)$.

In both examples, $\omega \mapsto \mathbb{E}(X|Z)(\omega)$ could be expressed as a function of $Z(\omega)$. This is no accident, but a consequence of the $\sigma(Z)$-measurability. The result that yields this statement is

(3.20) Lemma

Let X and Y be real RVs. Then

$$X \in m\sigma(Y) \quad \Leftrightarrow \exists f \in m\mathcal{B}(\mathbb{R}) : \forall \omega \in \Omega : X(\omega) = f(Y(\omega)).$$

Proof: '$\Leftarrow$' follows from 1.13 a).
'$\Rightarrow$':
(i): Assume first $X = 1_A$ with $A \in \mathcal{F}$. Since $X \in m\sigma(Y)$, we have $A \in \sigma(Y)$, and thus there exists $B \in \mathcal{B}(\mathbb{R})$ with $A = Y^{-1}(B)$. So, $X(\omega) = 1_B(Y(\omega))$.
(ii): Let $X = \sum_{i=1}^n \alpha_i 1_{A_i}$, with $A_i \cap A_j = \emptyset$ and $\alpha_i \neq \alpha_j$ if $i \neq j$. Then all A_i are in $\sigma(Y)$ (why?), and we find $B_i \in \mathcal{B}(\mathbb{R})$ to each A_i as above, and $X(\omega) = \sum_{i=1}^n \alpha_i 1_{B_i}(Y(\omega))$.
(iii) For $X \geqslant 0$, define

$$X_n(\omega) := 2^{-n} \lfloor 2^n X(\omega) \rfloor \wedge n.$$

By (ii), $X_n = f_n(Y)$ for some $f_n \in m\mathcal{B}(\mathbb{R})$ for all n. By (1.13 d), $x \mapsto f(x) := \sup_{n\in\mathbb{N}} f_n(x)$ is $\mathcal{B}(\mathbb{R})$-measurable, and

$$f(Y(\omega)) = \sup_n f_n(Y(\omega)) = \sup_n X_n(\omega) = X(\omega).$$

(iv): For general X, decompose into $X^+ - X^-$ and use (iii). □

(3.21) Remark, Definition, Statement

Can we extend (3.20) to RVs with values in arbitrary measurable spaces? Step (iii) in the proof seems to need the ordering of $\mathbb{R}$. But we can do much better (but not infinitely better):

Definition: A measurable space $(M, \mathcal{G})$ is called *(standard) Borel space* if it is isomorphic to a Borel-subset of $[0,1]$ as a measure space, i.e. if

$$\exists A \in \mathcal{B}([0,1]), f : M \to A \text{ bijective with } f \text{ and } f^{-1} \text{ measurable.}$$

This definition seems much more restrictive than it is. In particular, any separable, complete metric space ('Polish space') is Borel.

Statement: (3.20) holds if X maps into a Borel space, and Y into an arbitrary measurable space.
Proof: See Kallenberg, Foundations of Modern Probability (an excellent book!).

(3.22) Definition

Let X be an integrable real RV, and let $Z : \Omega \to \Omega'$ be any RV into some measurable space $(\Omega'\mathcal{F}')$. The map

$$\Omega' \to \mathbb{R}, \qquad z \mapsto \mathbb{E}(X|Z = z) := \mathbb{E}(X|Z)(\omega) \quad \text{with any } \omega \in Z^{-1}(z)$$

is called the conditional expectation of X given that $Z = z$. By (3.21) we know that this map is $(\mathcal{F}', \mathcal{B})$-measurable.

When X and Z are independent, we have the following nice formula.

(3.23) Proposition

Let $X : \Omega \to \Omega'$ and $Y : \Omega \to \Omega''$ be independent RVs, and let $h : \Omega' \times \Omega'' \to \mathbb{R}$ be $\mathcal{F}' \otimes \mathcal{F}''$-measurable and such that $h(X, Y)$ is integrable. Then

$$\mathbb{E}(h(X,Y)|Y)(\bar\omega) = \mathbb{E}(h(X, Y(\bar\omega))) \equiv \int h(X(\omega), Y(\bar\omega))\, \mathbb{P}(\mathrm{d}\omega).$$

Proof: Let first be $h(x, y) = 1_A(x)1_B(y)$ for $A \in \mathcal{F}'$, $B \in \mathcal{F}''$. Then

$$\mathbb{E}(h(X, Y(\bar\omega))) = \mathbb{E}(1_A(X)1_B(Y(\bar\omega))) = \mathbb{P}(X \in A)1_B(Y(\bar\omega)) \in m\sigma(Y), \qquad (*)$$

and for any $G \in \sigma(Y)$, we have

$$\begin{aligned}\int 1_G(\bar\omega)\mathbb{E}(h(X, Y(\bar\omega)))\mathbb{P}(\mathrm{d}\bar\omega) &= \mathbb{P}(X \in A) \int 1_G(\bar\omega) 1_B(Y(\bar\omega))\,\mathbb{P}(\mathrm{d}\bar\omega) \\ &\overset{\sigma(X)\perp\!\!\!\perp\sigma(Y)}{=} \mathbb{E}(1_A(X)1_B(Y)1_G) = \mathbb{E}(h(X,Y)1_G). \qquad (**)\end{aligned}$$

So in this special case, the claim holds. Since both $(*)$ and $(**)$ are stable under linear operations, the claim also holds if $h(x, y) = \sum_{i=1}^n 1_{A_i}(x)1_{B_i}(y)$ with $A_i \in \mathcal{F}'$ and $B_i \in \mathcal{F}''$. Since they are also stable under monotone limits, the claim holds for $h(x, y) = 1_C(x, y)$ for arbitrary $C \in \mathcal{F}' \otimes \mathcal{F}''$. Again by stability under linear operations, we can extend to $h = \sum_{i=1}^n \alpha_i 1_{C_i}$ with measurable C_i, by monotonicity to all nonnegative functions, and by taking positive and negative part to all integrable functions. □

(3.24) Example

Let (X_n) be iid integrable RVs, and $S_n = \sum_{i=1}^n X_i$. Let $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$. Then

$$\mathbb{E}(S_{n+1}|\mathcal{F}_n)(\omega) = S_n(\omega) + \mathbb{E}(X_{n+1}).$$

To prove this, use (3.23).

We finally look at another important special case of Definition (3.1):

(3.25) Definition

For $A \in \mathcal{F}$, $\mathcal{G} \subset \mathcal{F}$, the map

$$\omega \mapsto \mathbb{P}(A|\mathcal{G})(\omega) := \mathbb{E}(1_A|\mathcal{G})(\omega) \tag{3.1}$$

is called *conditional probability* of A given $\mathcal{G}$.

(3.26) Observation

We can think of the conditional probability as a map $(\omega, A) \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ of two variables. For each fixed $A \in \mathcal{F}$, equation (3.1) defines a $\mathcal{G}$-measurable RV. For each *fixed* sequence $(A_n) \subset \mathcal{F}$ with $A_n \cap A_m = \emptyset$ for $m \neq n$, monotone convergence guarantees that

$$\mathbb{P}(\bigcup_{n=1}^\infty A_n|\mathcal{G}) = \sum_{n=1}^\infty \mathbb{P}(A_n|\mathcal{G}) \qquad \text{almost surely.}$$

Thus, for all ω from a set Ω_1 of measure 1, the map $A \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ is σ-additive for the members of that fixed sequence. It is tempting to think of the map $A \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ as a probability measure, but the problem is that for this we need σ-additivity for all sequences of mutually disjoint sets. But since for each sequence we could get a different set of measure zero where σ-additivity fails, and since there are uncountably many sequences of disjoint sets, we can not in general find a common set of measure 1 where σ-additivity holds. We solve this problem in the mathematicians way.

(3.27) Definition

Let $\mathcal{G} \subset \mathcal{F}$. A map $\mu : \Omega \times \mathcal{F} \to [0,1]$ with the properties that
(i): $\forall A \in \mathcal{F} : \omega \mapsto \mu(\omega, A)$ is measurable, and $\mu(\omega, A) = \mathbb{P}(A|\mathcal{G})(\omega)$ almost surely,
(ii): For all $\omega \in \Omega$, $A \mapsto \mu(\omega, A)$ is a probability measure
is called *regular conditional probability* of $\mathbb{P}$ given $\mathcal{G}$.

(3.28) Example

In the situation of Example (3.19), the map

$$(\omega, A) \mapsto \int_A f_{X|Z}(x; Z(\omega))\,\mathrm{d}x$$

is a regular conditional probability of $\mathbb{P}_X$ given $\mathcal{G} = \sigma(Z)$.

In more general situations, we cannot construct regular conditional probabilities so explicitly. But we have

(3.29) Theorem

Let $(\Omega, \mathcal{F})$ be a standard Borel space (see (3.21)). Then for each σ-algebra $\mathcal{G} \subset \mathcal{F}$ and all probability measures $\mathbb{P}$ on $(\Omega, \mathcal{F})$, the regular conditional probability of $\mathbb{P}$ given $\mathcal{G}$ exists.

Proof: We assume first that $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For each $r \in \mathbb{Q}$, let

$$\omega \mapsto F(r, \omega) = \mathbb{P}((-\infty, r]|\mathcal{G})(\omega)$$

be a conditional probability of $(-\infty, r]$ given $\mathcal{G}$. Since $r \mapsto 1_{(-\infty,r]}$ is a growing set function, monotonicity of conditional expectation guarantees that, for $r < s$, we find a set $A_{r,s} \in \mathcal{F}$ with $\mathbb{P}(A_{r,s}) = 1$ and,

$$F(r, \omega) = \mathbb{E}(1_{(-\infty,r]}|\mathcal{G})(\omega) \leqslant \mathbb{E}(1_{(-\infty,s]}|\mathcal{G})(\omega) = F(s, \omega) \qquad \forall \omega \in A_{r,s}.$$

By dominated convergence, we also find, for every $r \in \mathbb{Q}$, a set B_r with $\mathbb{P}(B_r) = 1$ and

$$\lim_{n\to\infty} F(r + 1/n, \omega) = F(r, \omega) \qquad \forall \omega \in B_r.$$

Again by dominated convergence, we find a set $C \in \mathcal{F}$ with $\mathbb{P}(C) = 1$ and

$$0 = \inf_{n\in\mathbb{N}} F(-n, \omega) = 1 - \sup_{n\in\mathbb{N}} F(n, \omega) \qquad \forall \omega \in C.$$

So for all $\omega \in \Omega_0 := \bigcap_{r,s\in\mathbb{Q}} A_{r,s} \cap \bigcap_{r\in\mathbb{Q}} B_r \cap C$ (with $\mathbb{P}(\Omega_0) = 1$), we define

$$\tilde{F}(z, \omega) := \inf_{r\in\mathbb{Q}, r>z} F(r, \omega) \qquad \forall z \in \mathbb{R}.$$

By construction and our above observations, the map $z \mapsto \tilde{F}(z, \omega)$ is a distribution function. Let F_0 be an arbitrary fixed distribution function and set $\tilde{F}(z, \omega) = F_0(z)$ for all $\omega \in \Omega_0$. Let $\mu(\omega, .)$ be the probability measure on $\mathbb{R}$ induced by the distribution function $F(., \omega)$. Then the map

$$\omega \mapsto \mu(\omega, (-\infty, r]) = \tilde{F}(r, \omega) 1_{\Omega_0}(\omega) + F_0(r) 1_{\Omega_0^c}(\omega)$$

is measurable for each fixed r, and since the intervals $(-\infty, r]$, $r \in \mathbb{R}$, are a $\cap$-stable generator of $\mathcal{B}(\mathbb{R})$, the map $\omega \mapsto \mu(\omega, A)$ is measurable for all $A \in \mathbb{R}$. By construction, $A \mapsto \mu(\omega, A)$ is also a measure for each fixed $\omega \in \Omega$.

For all $\omega \in \Omega_0$, by definition $\mu(\omega, A) = \mathbb{P}(A|\mathcal{G})(\omega)$ for all A of the form $A = (-\infty, r]$, $r \in \mathbb{Q}$. Again by extension of a $\cap$-stable generator, this holds for all $A \in \mathcal{B}(\mathbb{R})$. So, $A \mapsto \mu(\omega, A)$ is a conditional probability of given $\mathcal{G}$ almost surely, and we have proved the theorem in the case $\Omega = \mathbb{R}$.

In the general case, let $(\Omega, \mathcal{F})$ be a Borel space and $\mathbb{P}$ be a probability measure on it, and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Then there exists a $B \in \mathcal{B}$ and an isomorphism of measurable spaces $\varphi : (\Omega, \mathcal{F}) \to (B, B \cap \mathcal{B}(\mathbb{R}))$. Then $\tilde{\mathbb{P}} := \mathbb{P} \circ \varphi^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ (supported on B), and $\tilde{\mathcal{G}} = \varphi(\mathcal{G})$ is a sub-σ-algebra of $\mathcal{B}(\mathbb{R}) \cap B$. Let $(x, \tilde{A}) \mapsto \tilde{\mu}(x, \tilde{A})$ be the regular conditional probability measure of $\tilde{\mathbb{P}}$ given $\tilde{\mathcal{G}}$. Then you can check that $\mu(\omega, A) := \tilde{\mu}(\varphi(\omega), \varphi(A))$ is a regular conditional probability of $\mathbb{P}$ given $\mathcal{G}$. □

In the final example of this section, the conditional expectation can be computed explicitly, but it is not completely easy to do so.

(3.30) Example

Let (X_i) be iid integrable RVs, and let $S_n = \sum_{i=1}^n X_i$. Let $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \ldots)$. Show as an exercise that

$$\mathbb{E}(X_1|\mathcal{G}_n) = \mathbb{E}(X_2|\mathcal{G}_n) = \ldots = \mathbb{E}(X_n|\mathcal{G}_n) = \frac{1}{n} S_n.$$

Do this using the following steps:

1) Show that $\mathcal{G}_n = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$.

2) Show that $\mathbb{E}(X_i|\mathcal{G}_n) = \mathbb{E}(X_i|S_n)$ for all $i \leqslant n$.

3) Show that $\mathbb{E}(X_i|S_n) = \mathbb{E}(X_j|S_n)$ almost surely for $i, j \leqslant n$.

4) Conclude the claim.

The statement says that if we sum n independent RVs, and if we know that the sum has the value M, then the best guess for each of them is simply M/n.

4. Martingales

(4.1) Motivating Example

We consider a simple game of chance: in each round, a coin is thrown, and you have to decide the amount a that you want to bet. If the coin comes up heads, you win the amount a. Otherwise, you lose the amount a.

We model this: Let (X_i) be iid RVs with $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$. Your expected win in round i is exactly $\mathbb{E}(X_i) = 0$. Let $(a_i(x_1, \ldots, x_{i-1}))_{i\in\mathbb{N}}$ be your gambling strategy: based on the outcomes of the coins, 1 to $i-1$, you decide to bet the amount a_i. Then

$$Y_n(\omega) := a_n(X_1(\omega), \ldots, X_{n-1}(\omega))X_n$$

is your win/loss in round n, and

$$S_n(\omega) = \sum_{i=1}^{n} Y_i(\omega)$$

is your total win/loss after n games.

a) Even though the X_i are independent, the Y_i usually are not independent at all. On the other hand, your strategy can only depend on the past for obvious reasons. So at least the Y_i are independent of all the X_{i+j}, $j > 0$.

b) Observe that this strategy includes the famous doubling strategy in Roulette (except that in Roulette there is the zero): start with $a_1 = 1$ and double every time you lose, until you win, say that this happens in game j. It is easy to see that at this point you have won exactly one unit. Then start again at $a_{j+1} = 1$ and repeat. Many intelligent people think that one can win Roulette like this.

c) Let's calculate

$$\mathbb{E}(S_n) = \sum_{j=1}^{n} \mathbb{E}(a_j(X_1, \ldots, X_{j-1})X_j) = \sum_{j=1}^{n} \mathbb{E}(a_j(X_1, \ldots, X_{j-1}))\mathbb{E}(X_j) = 0.$$

So, our expected gain is exactly zero, no matter what strategy we try!

d) Assume that after j games, we have observed the outcomes $X_1, \ldots, X_j$. What is our expected gain after $j+1$ games, given these outcomes? We set $\mathcal{F}_j = \sigma(X_1, \ldots, X_j)$ and calculate

$$\mathbb{E}(S_{j+1}|\mathcal{F}_j) = \sum_{i=1}^{j} Y_i + \mathbb{E}(a_{j+1}(X_1, \ldots, X_j)X_{j+1}|\mathcal{F}_j) = S_j + a_{j+1}(X_1, \ldots, X_j)\mathbb{E}(X_{j+1}) = S_j.$$

So, at any point during the game, on average we stay exactly as rich as we are right at that point.

e) All considerations still work when the X_i are arbitrary independent RVs with $\mathbb{E}(X_i) = 0$.

f) The outcome of all this is: in *fair games* like the above one, there is *no winning strategy.*

g) Real games are more complicated: The game you play in the next round (i.e. the distribution of the RV X_{i+1}) may depend on the outcome of all the previous rounds. Your strategy may be based on more information than just the outcome of the previous games (e.g. Poker: the face of your opponent).

h) Martingales are those games where despite all this, there still is no winning strategy. On top of that, they are one of the most powerful concepts of modern probability theory. The main reason for this is that they allow to consider RVs that are not at all independent, but at the same time still get very powerful general results.

(4.2) Definition

a) A collection $(X_n)_{n\in\mathbb{N}}$ of real RVs is called a (discrete time) *stochastic process* (SP).

b) Let $(X_n)_{n\in\mathbb{N}}$ be a stochastic process. A probability measure μ on $\Omega = \mathbb{R}^{\mathbb{N}}$, with $\mathcal{F} = \mathcal{B}(\mathbb{R})^{\otimes\mathbb{N}}$ (product σ-algebra), is called the canonical representation of $(X_n)_{n\in\mathbb{N}}$ if for all $n \in \mathbb{N}$ and all $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mu(A_1 \times \ldots \times A_n \times \mathbb{R}^{\mathbb{N}}).$$

(4.3) Theorem

For an arbitrary stochastic process (X_n), the canonical representation exists and is unique (as a measure).

Proof: exercise; see also Kallenberg or Durrett. In the course 'Stochastic Processes' we will see a much more powerful existence result (Kolmogorovs extension theorem). □

(4.4) Definition

a) A family $(\mathcal{F}_n)$ of σ-algebras is called a *filtration* if for each $n \in \mathbb{N}$, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

b) A stochastic process (X_n) is called *adapted to a filtration* $(\mathcal{F}_n)$ if $X_n \in m\mathcal{F}_n$ for all n.

(4.5) Examples

a) Let (X_n) be a SP, $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$. Then (X_n) is $(\mathcal{F}_n)$-adapted, and $Y \in \mathcal{F}_n \Leftrightarrow Y = f(X_1, \dots, X_n)$ for some measurable $f : \mathbb{R}^n \to \mathbb{R}$ (by (3.20)).

b) $\sigma(X_1, \dots, X_n)$ is the minimal filtration such that (X_n) is $(\mathcal{F}_n)$-adapted (why? Why not $\mathcal{F}_n = \sigma(X_n)$?). But $(\mathcal{F}_n)$ can be much larger than that. Find simple examples for larger filtrations!

c) In Example (4.1), with $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$, we have that (Y_n) is $(\mathcal{F}_n)$-adapted, and that (a_n) is $\mathcal{F}_{n-1}$-adapted.

(4.6) Definition

Let $(\mathcal{F}_n)$ be a filtration and (X_n) an $(\mathcal{F}_n)$-adapted process with $\mathbb{E}(|X_n|) < \infty$ for all n. (X_n) is called a

a) *martingale* if $\mathbb{E}(X_{n+1}|\mathcal{F}_n)(\omega) = X_n(\omega)$ for almost all ω.

b) *supermartingale* if $\mathbb{E}(X_{n+1}|\mathcal{F}_n)(\omega) \leqslant X_n(\omega)$ for almost all ω.

c) *submartingale* if $\mathbb{E}(X_{n+1}|\mathcal{F}_n)(\omega) \geqslant X_n(\omega)$ for almost all ω.

(4.7) Remark

a) (X_n) submartingale $\Leftrightarrow$ $(-X_n)$ supermartingale.

b) (X_n) and $(-X_n)$ submartingales $\Leftrightarrow$ (X_n) is a martingale.

c) Let $(X_n)_{n \geqslant 0}$ be a stochastic process. Then:
(X_n) is a (sub-), (super-) martingale $\Leftrightarrow$ $(X_n - X_0)$ is a (sub-), (super-) martingale.

d) These remarks allow us to state many facts only for submartingales: They then also hold for martingales (by b) $\Rightarrow$), and the 'negative' statements hold for supermartingales (by a)). c) says that (de-)randomizing the starting point of a SP does not change its martingale property.

(4.8) Proposition

Let (X_n) be a submartingale. Then

$$\forall m \geqslant 0 : \mathbb{E}(X_{n+m}|\mathcal{F}_n) \geqslant X_n \text{ a.s.}$$

Proof: Exercise. □

(4.9) Examples

a) X_n independent integrable RVs, $\mathcal{F}_n = \sigma(X_i : i \leqslant n)$, $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Then

$$S_n := \sum_{i=1}^{n} X_i \qquad \text{is a} \begin{cases} \text{martingale} & \text{if } \forall n : \mathbb{E}(X_n) = 0, \\ \text{submartingale} & \text{if } \forall n : \mathbb{E}(X_n) \geqslant 0, \\ \text{supermartingale} & \text{if } \forall n : \mathbb{E}(X_n) \leqslant 0. \end{cases}$$

b) X_n independent RVs, $\mathbb{E}(X_n) = 1$ for all n, $\mathcal{F}_n$ as in a). Put $M_0 = 1$, $M_n = \prod_{i=1}^n X_i$. Then (M_n) is a martingale (exercise!). In particular, this is true for $M_n = \exp(\sum_{j=1}^n Y_j)$ when $\mathbb{E}(\exp Y_j) = 1$ and the Y_j are independent.

c) Let X be an integrable RV, $(\mathcal{F}_n)$ an arbitrary filtration. Put $Y_n := \mathbb{E}(X|\mathcal{F}_n)$ (information available on X by only considering sets from $\mathcal{F}_n$). Then

$$\mathbb{E}(Y_n|\mathcal{F}_{n-1}) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_n)|\mathcal{F}_{n-1}) = \mathbb{E}(X|\mathcal{F}_{n-1}) = Y_{n-1},$$

so (Y_n) is a martingale.

We now extend the idea of a gambling strategy from (4.1).

(4.10) Definition

Let $(\mathcal{F}_n)$ be a filtration. A stochastic process (C_n) is called *previsible* if $C_n \in m\mathcal{F}_{n-1}$ for all n.

(4.11) Defintion

Let $(\mathcal{F}_n)$ be a filtration, (X_n) an $(\mathcal{F}_n)$-adapted SP, and (C_n) an $(\mathcal{F}_n)$-previsible process. The process (Y_n) with

$$Y_n(\omega) := \sum_{k=1}^{n} C_k(\omega)(X_k(\omega) - X_{k-1}(\omega))$$

is called the *discrete stochastic integral* of (C_n) with integrator (X_n).

In this case, we write $Y_n = (C \bullet X)_n$.

If (X_n) is a martingale, (Y_n) is called the *martingale transform* of (X_n) by (C_n).

(4.12) Theorem

Let $(\mathcal{F}_n)$ be a filtration, (C_n) previsible, (X_n) a submartingale (supermartingale). Assume either (i) that $\|C_n\|_\infty < \infty$ for all n,
or (ii) that $C_n \in L^2$ and $X_n \in L^2$.

Then

a) If $C_n \geqslant 0$ for all n, then $(C \bullet X)_n$ is a submartingale (supermartingale).

b) If (X_n) is a martingale, then also $(C \bullet X)_n$ is a martingale.

Proof: Both (i) and (ii) guarantee that $\mathbb{E}(|(C \bullet X)_n|) < \infty$ for all n. Furthermore,

$$\mathbb{E}((C \bullet X)_n | \mathcal{F}_{n-1}) = (C \bullet X)_{n-1} + \mathbb{E}((C \bullet X)_n - (C \bullet X)_{n-1} | \mathcal{F}_{n-1}) =$$
$$= (C \bullet X)_{n-1} + \mathbb{E}(C_n(X_n - X_{n-1}) | \mathcal{F}_{n-1}) = (C \bullet X)_{n-1} + C_n\Big(\mathbb{E}(X_n | \mathcal{F}_{n-1}) - X_{n-1}\Big).$$

The two claims now follow immediately. □

(4.13) Remark

The interpretation of Theorem 4.12 is maybe more important than the statement itself.

1) We think of each $\omega \in \Omega$ as one of the (pre-determined, but unknown) possible series of outcomes of all rounds of a game of chance.

2) The process (X_n) is the amount of money you have in round n if you win/lose precisely $(X_j - X_{j-1})$ Euro in round j, $j \leqslant n$, and if the game is governed by ω.

3) The σ-algebra $\mathcal{F}_{n-1}$ is all the information that you have in round $n-1$; this can be just the outcomes of the previous rounds (then $\mathcal{F}_n = \sigma(X_j : j \leqslant n)$), or it can contain additional information on the previous rounds, such as the face of your opponent after getting her hand of cards for the current round.

4) If (X_n) is a martingale, then $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$ means that if we have perfect information

on everything that can be known up to round $n-1$, based on that information the best guess for our money after round n is to be exactly equal to what we have in round $n-1$.

5) For a supermartingale, it is as in 4), but now the best guess is that we will lose money in round n (unfavourably unfair game).

6) The fact that the process (C_n) is previsible means that the value of $C_n(\omega)$ depends entirely on the information up to round $n-1$. So, C_n is a gambling strategy for round n that is allowed to use all information from the past, but cannot use information for the future. This means that in round n, we bet exactly C_n Euro.

7) The process $(C \bullet X)_n$ is the money that you have after round n if in round $j \leqslant n$, you increase/decrease all winnings/losings by a factor C_j.

8) The statement of (4.12) says that *you can't beat the system*: Statement a) says that if a game is rigged against you, and if you can only bet positive amounts of money, then your best strategy is to stop playing immediately. Everything else will lose you money more likely than not. Statement b) says that if a game is fair, then no strategy will make it possible to win (or lose) money of average.

A very important special gambling strategy are stopping times.

(4.14) Definition

Let $(\mathcal{F}_n)$ be a filtration. A RV $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is called a $(\mathcal{F}_n)$-*stopping time* (or simply stopping time) if

$$\forall n \in \mathbb{N}_0 \cup \{\infty\} : \{T \leqslant n\} \in \mathcal{F}_n.$$

(4.15) Lemma

T is an $(\mathcal{F}_n)$-stopping time if and only if $\{T = n\} \in \mathcal{F}_n$ for all n.

Proof: Exercise. □

(4.16) Examples

$(X_n)_{n\in\mathbb{N}}$ SP, $\mathcal{F}_n = \sigma(X_j : j \leqslant n)$.

a) $T(\omega) := \inf\{k \in \mathbb{N} : X_k(\omega) \geqslant c\}$ is a stopping time, since

$$\{T > n\} = \{\forall k \leqslant n : X_k < c\} = \bigcap_{k=1}^{n} \underbrace{X_k^{-1}((-\infty, c])}_{\in \mathcal{F}_k} \in \mathcal{F}_n.$$

b) $T := \inf\{n : X_{n+4} \geqslant c\}$ is not an $(\mathcal{F}_n)$-stopping time, but it is an $(\mathcal{F}_{n+4})$-stopping time.

c) $S := \sup\{n : X_n \geqslant c\}$ is usually not a stopping time. Exercise: find natural examples where S is not a stopping time, and find special cases where S is a stopping time.

(4.17) Definition

Let (X_n) be an $(\mathcal{F}_n)$-adapted SP and T be an $(\mathcal{F}_n)$-stopping time. The SP $(X_{T\wedge n})_n$ with $X_{T\wedge n}(\omega) := (X_{T(\omega)\wedge n})(\omega)$ is called the process (X_n) stopped at T. Sometimes this definition is extended to cases where T is not a stopping time.

(4.18) Theorem

Let (X_n) be an $(\mathcal{F}_n)$-adapted process, $X_0 \equiv 0$ and T be an $(\mathcal{F}_n)$-stopping time. Then:

a) $(X_{T\wedge n})$ is $(\mathcal{F}_n)$-adapted.

b) If (X_n) is a (sub-, super-)martingale, then $(X_{T\wedge n})$ also is a (sub-, super-)martingale.

Proof:

$$X_{T\wedge n}(\omega) = \sum_{i=1}^{n} C_i(\omega)(X_i(\omega) - X_{i-1}(\omega)) \qquad \text{with } C_i(\omega) := \begin{cases} 0 & \text{if } T(\omega) < i \\ 1 & \text{if } T(\omega) \geqslant i. \end{cases}$$

(To see this, check the cases $\{T < n\}$ and $\{T \geqslant n\}$ separately.) Since $\{C_n = 0\} = \{T < n\} = \{T \leqslant n-1\} \in \mathcal{F}_{n-1}$, the process (C_n) is previsible. In particular, a) holds, and b) directly follows from Theorem (4.12) (i) since $|C_j| \leqslant 1$ for all j. □

(4.19) Example: Distribution is Random Walk Hitting Times

Let $(X_n)_{n\in\mathbb{N}}$ be iid with $\mathbb{P}(X_i = \pm 1) = 1/2$, $X_0 \equiv 0$. Set $S_n = \sum_{j=1}^{n} X_j$ (simple random walk), $\mathcal{F}_n := \sigma(X_j : j \leqslant n) \overset{\text{check this!}}{=} \sigma(S_j : j \leqslant n)$, and $T := \inf\{n : S_n \geqslant 1\} \leqslant \infty$. Since (S_n) is $(\mathcal{F}_n)$-adapted, T is a stopping time. Can we compute its distribution? I.e., can we find out how long it will take until the random walk is positive for the first time?

To do it, fix $\theta > 0$ and let

$$M_n(\omega) := M_{\theta,n}(\omega) = \frac{1}{(\cosh\theta)^n}\,\mathrm{e}^{\theta S_n(\omega)} = \prod_{i=1}^{n}\Big(\frac{1}{\cosh\theta}\,\mathrm{e}^{\theta X_i(\omega)}\Big), \qquad M_0 := M_{\theta,0} := 1.$$

Since $\mathbb{E}(\mathrm{e}^{\theta X_i}) = \frac{1}{2}(\mathrm{e}^{\theta} + \mathrm{e}^{-\theta}) = \cosh(\theta)$, (M_n) is a martingale by Example (4.9 b), and thus $\mathbb{E}(M_n) = \mathbb{E}(M_0) = 1$. By (4.18 b), also $\mathbb{E}(M_{T\wedge n}) = 1$ for all n. Next we show that $M_{T\wedge n}$ converges pointwise: we have that

$$\frac{M_{T(\omega)\wedge n+1}(\omega)}{M_{T(\omega)\wedge n}(\omega)} = \begin{cases} 1 & \text{if } T(\omega) \leqslant n, \\ \mathrm{e}^{\theta} / \cosh\theta & \text{if } T(\omega) = n+1, \\ \mathrm{e}^{\theta X_{n+1}(\omega)} / \cosh\theta \leqslant \mathrm{e}^{\theta} / \cosh\theta & \text{if } T(\omega) > n+1. \end{cases}$$

The second case holds since the step from 0 to 1 must be positive. We conclude that

$$\lim_{n\to\infty} M_{T(\omega)\wedge n}(\omega) = M_{T(\omega)}(\omega) 1_{\{T(\omega)<\infty\}}$$

exists for all $\omega \in \Omega$. (This is a special case of the famous Martingale convergence theorem we will prove later.) We can even compute the pointwise limit in case of $T(\omega) < \infty$,

$$M_{T(\omega)}(\omega) = (\cosh\theta)^{-T(\omega)}\, \mathrm{e}^{\theta S_{T(\omega)}} = (\cosh\theta)^{-T(\omega)}\, \mathrm{e}^{\theta}\,.$$

Now we can apply dominated convergence, since $S_{T\wedge n} \leqslant 1$:

$$1 = \lim_{n\to\infty} \mathbb{E}(M_{T\wedge n}) = \mathbb{E}(\lim_{n\to\infty} M_{T\wedge n}) = \mathrm{e}^{\theta}\, \mathbb{E}((\cosh\theta)^{-T} 1_{\{T<\infty\}}).$$

This holds for all $\theta > 0$. Taking the limit $\theta \to 0$ (dominated convergence again!), we find first that $\mathbb{P}(T < \infty) = 1$. Using this, and setting $\alpha := 1/\cosh\theta$, we find that

$$\mathrm{e}^{-\theta} = \mathbb{E}(\alpha^T) = \sum_{j=1}^{\infty} \alpha^j \mathbb{P}(T = j),$$

i.e. we have computed the probability generating function of T. Therefore, and since $\mathrm{e}^{-\theta} = \cosh\theta - \sinh\theta = \frac{1-\sqrt{1-\alpha^2}}{\alpha}$ by the addition theorems for cosh and sinh, we conclude that

$$\mathbb{P}(T = j) = \frac{\mathrm{d}^j}{\mathrm{d}\alpha^j} \frac{1-\sqrt{1-\alpha^2}}{\alpha}|_{\alpha=0} = (-1)^{m+1} \frac{\Gamma(3/2)}{\Gamma(m+1)\Gamma(3/2-m)} 1_{\{j=2m-1\}}.$$

This decays like $m^{-3/2}$ for large m.

(4.20) Observation

In (4.19), in the context of gambling T is the strategy to go home after winning exactly one Euro. Since $\mathbb{P}(T < \infty) = 1$ and $S_T = 1$ on $\{T < \infty\}$, we have $\mathbb{E}(S_T) = 1$. So, with this strategy, we seem to win 1 Euro on average, despite the fact that S_n is a martingale.

This seems to be a contradiction with (4.13). Can we beat the system after all?

The answer is no. In practice, we have to stop playing at some point, and $\mathbb{E}(S_{T\wedge n}) = 0$ for all n. Picturing the different paths of the random walk, there are many paths where for large n, $T \leqslant n$ holds, and those contribute to our winnings. But for the very few paths where still $T > n$, the random walk S_n tends to be very far away from zero; i.e. the probability that we are losing is small, but if we are losing, we are losing a lot. If we increase n further, this split becomes more extreme: more paths will give us winnings, but the losses on the losing paths will be even more terrible.

The limit $n \to \infty$ destroys all losing scenarios, but for finite n they exactly balance the winning ones. Mathematically, what happens is simply that we cannot exchange limit and expectation in the expression $0 = \lim_{n\to\infty} \mathbb{E}(S_{T\wedge n})$.

Our next Theorem investigates conditions under which such an exchange of limits is allowed. Note that by our calculations in (4.19), $\mathbb{E}(T) = \infty$.

(4.21) Doobs Optional Stopping Theorem

Let (X_n) be an $(\mathcal{F}_n)$-submartingale, and T an $(\mathcal{F}_n)$-stopping time. Assume at least one of the following conditions:

(i): T is a.s. bounded, i.e. $\exists N \in \mathbb{N} : \mathbb{P}(T \geqslant N) = 0$.
(ii): (X_n) is a.s. uniformly bounded and T is a.s. finite, i.e.

$$\exists K < \infty : \forall n \in \mathbb{N} : \mathbb{P}(|X_n| > K) = 0, \qquad \text{and } \mathbb{P}(T = \infty) = 0.$$

(iii): T is integrable and (X_n) has a.s. bounded increments, i.e.

$$\mathbb{E}(T) < \infty \qquad \text{and} \qquad \exists K < \infty : \forall j \in \mathbb{N} : \mathbb{P}(|X_{j+1} - X_j| > K) = 0.$$

Then

$$\mathbb{E}(X_T) \geqslant \mathbb{E}(X_0).$$

Proof: By (4.18), $(X_{T\wedge n})$ is a submartingale.

If (i) holds, then $\mathbb{E}(X_T - X_0) = \mathbb{E}(X_{T\wedge N} - X_0) \geqslant 0$.

If (ii) holds, then for all n, $|X_{T\wedge n} - X_0| \leqslant 2K$ a.s., and by dominated convergence then

$$\mathbb{E}(X_T - X_0) = \lim_{n\to\infty} \mathbb{E}(X_{T\wedge n} - X_0) \geqslant 0.$$

If (iii) holds, then almost surely for all n,

$$|X_{T(\omega)\wedge n}(\omega) - X_0(\omega)| = \Big| \sum_{k=1}^{T(\omega)\wedge n} (X_k(\omega) - X_{k-1}(\omega)) \Big| \leqslant KT(\omega)$$

Since T is integrable by assumption, dominated convergence as above gives the result. □

(4.22) Corollary

If in (4.21), (X_n) is a martingale (supermartingale), then under the same assumptions $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ $(\mathbb{E}(X_T \leqslant \mathbb{E}(X_0))$.

(4.23) Corollary

Let (M_n) be a martingale with bounded increments, (C_n) a uniformly bounded previsible process, and T an integrable stopping time, all with respect to some filtration $(\mathcal{F}_n)$. Then $\mathbb{E}((C \bullet M)_T) = 0$.

Proof: $C \bullet M$ is a martingale with bounded increments. Apply (4.22). □

A good sufficient condition for $\mathbb{E}(T) < \infty$ is given below. It says that at any time n, conditional on the past, the stopping happens in the next N steps with probability at least ε.

(4.24) Lemma

Let $(\mathcal{F}_n)$ be a filtration with $\mathcal{F}_0 = \{\emptyset, \Omega\}$, T be an $(\mathcal{F}_n)$-stopping time, and assume

$$\exists N \in \mathbb{N}, \varepsilon > 0 : \forall n \in \mathbb{N}_0 : \mathbb{P}(T \leqslant N + n | \mathcal{F}_n) > \varepsilon \text{ a. s.}$$

Then $\mathbb{E}(T) < \infty$.

Proof: We first show by induction that

$$(*) \qquad \mathbb{P}(T > kN) < (1-\varepsilon)^k \qquad \forall k \in \mathbb{N}.$$

For $k = 1$, this holds since

$$\mathbb{P}(T > N) = \mathbb{P}(T > N|\mathcal{F}_0) = 1 - \mathbb{P}(T \leqslant N|\mathcal{F}_0) < 1 - \varepsilon.$$

Assuming that we have shown $(*)$ up to some $k \in \mathbb{N}$, we have

$$\begin{aligned} &\mathbb{P}(T > (k+1)N) = \mathbb{E}(\mathbb{E}(1_{\{T>(k+1)N\}} 1_{\{T>kN\}} |\mathcal{F}_{kN})) = \\ &= \mathbb{E}(1_{\{T>kN\}} \quad \underbrace{\mathbb{E}(1_{\{T>(k+1)N\}}|\mathcal{F}_{kN})}_{<1-\varepsilon \text{ by assumption, with } n=kN} \quad) < (1-\varepsilon)\mathbb{P}(T > kN) < (1-\varepsilon)^{k+1}. \end{aligned}$$

Therefore,

$$\mathbb{E}(T) = \sum_{n=1}^{\infty} \mathbb{P}(T \geqslant n) \leqslant N \sum_{k=0}^{\infty} \mathbb{P}(T > kN) = \frac{N}{\varepsilon} < \infty.$$

□

The next theorem is one of the most important results of martingale theory.

(4.25) Doobs Martingale Convergence Theorem

Let (X_n) be a supermartingale which is uniformly bounded in L^1, i.e. with $\sup\{\mathbb{E}(|X_j|) : j \in \mathbb{N}\} < \infty$. Then $\lim_{n\to\infty} X_n$ exists and is finite almost surely.

For the proof, we need some preparations.

(4.26) A gambling strategy

Let (X_n) be a SP. As before, $X_n - X_{n-1}$ represents the win/loss in round n. X_n is the total win/loss that you would get after round n by playing the game every round, but you can choose to not play in some rounds. Consider the following strategy:

1) Pick two numbers $a < b$.

2) Do not play the game until $X_{n_1}(\omega) < a$ for some n_1, then start playing until $X_{n_2}(\omega) > b$ for some $n_2 > n_1$.

3) Stop playing at time n_2, until a time $n_3 > n_2$ when again $X_{n_3}(\omega) < a$, then start playing again until some n_4 for which $X_{n_4}(\omega) > b$.

4) Continue in this way indefinitely. Each time between n_{2j-1} and n_{2j} ($j \in \mathbb{N}$), we win at least $b - a$.

Formally, we set $C_1 = 1_{\{X_0<a\}}$, and

$$C_n = 1_{\{C_{n-1}=1\}} 1_{\{X_{n-1} \leqslant b\}} + 1_{\{C_{n-1}=0\}} 1_{\{X_{n-1}<a\}}.$$

(C_n) is previsible and implements the above gambling strategy. The main idea of the proof of Theorem (4.25) simply is that if (X_n) is a supermartingale (or a martingale), then this strategy cannot be successful, since we cannot beat the system. Therefore for almost all $\omega \in \Omega$ and all intervals $[a, b]$, the sequence $n \mapsto X_n(\omega)$ only 'crosses' $[a, b]$ finitely many times. We will now see that this idea actually works.

(4.27) Definition

Let $\boldsymbol{x} = (x_n)_{n \in \mathbb{N}}$ be any real-valued sequence. For $a < b$ let

$$u_{N,[a,b]}(\boldsymbol{x}) := \max\{k \in \mathbb{N} : \exists s_1 < t_1 < s_2 < t_2 < \ldots < s_k < t_k < N \text{ with } x_{s_i} < a, x_{t_i} > b \,\forall i\}$$

denote the number of upcrossings of the interval $[a, b]$ performed by $\boldsymbol{x}$ before time N.

(4.28) Lemma

Let $\boldsymbol{x}$ be any real sequence, and assume that for all $a, b \in \mathbb{Q}$, $a < b$, we have

$$\lim_{N \to \infty} u_{N,[a,b]}(\boldsymbol{x}) < \infty.$$

Then $\lim_{n \to \infty} x_n$ exists in $[-\infty, \infty]$.

Proof: Assume that the limit does not exist. Then we can find $a, b \in \mathbb{Q}$ with $\liminf x_n < a < b < \limsup x_n$, which means that $\lim_{N \to \infty} u_{N,[a,b]}(\boldsymbol{x}) = \infty$ for these a, b. This is excluded by assumption, so (x_n) must converge. □

(4.29) Doobs Upcrossing Lemma

Let $\boldsymbol{X} = (X_n)$ be a supermartingale, $a < b$,

$$C_n = 1_{\{C_{n-1}=1\}} 1_{\{X_{n-1} \leqslant b\}} + 1_{\{C_{n-1}=0\}} 1_{\{X_{n-1} < a\}},$$

and $Y_n = (C \bullet X)_n$. Define

$$\mathcal{U}_{N,[a,b]}(\omega) := u_{N,[a,b]}(\boldsymbol{X}(\omega)).$$

Then

a) For all $N \in \mathbb{N}$ and all $\omega \in \Omega$,

$$Y_N(\omega) \geqslant (b - a)\mathcal{U}_{N,[a,b]}(\omega) - \max\{a - X_N(\omega), 0\}.$$

b) $(b - a)\mathbb{E}(\mathcal{U}_{N,[a,b]}) \leqslant \mathbb{E}(\max\{a - X_N, 0\})$.

Proof: (i) is clear: each crossing of $[a, b]$ contributes at least $b - a$, and in case that $C_N(\omega) = 1$, the loss in the last unfinished upcrossing period is not larger than $|a - X_N(\omega)|$.
(ii): Since (C_n) is previsible and bounded, (Y_n) is a supermartingale. Thus $\mathbb{E}(Y_N) \leqslant \mathbb{E}(Y_0) = 0$. The result follows by integrating the inequality in a) and rearranging. □

(4.30) Corollary

In the situation of (4.29), assume in addition that $c := \sup_n \mathbb{E}(|X_n|) < \infty$. Then

$$\mathbb{P}(\lim_{N \to \infty} \mathcal{U}_{N,[a,b]} = \infty) = 0.$$

Proof: By (4.29) and $\max\{a-b,0\} \leqslant |a|+|b|$, we have

$$(b-a)\mathbb{E}(\mathcal{U}_{N,[a,b]}) \leqslant |a| + \mathbb{E}(|X_N|) \leqslant |a| + c.$$

Since $N \mapsto \mathcal{U}_{N,[a,b]}$ is monotone increasing, monotone convergence gives

$$\mathbb{E}(\lim_{N\to\infty} \mathcal{U}_{N,[a,b]}) \leqslant \frac{|a|+c}{b-a} < \infty,$$

and thus $\mathbb{P}(\lim_{N\to\infty} \mathcal{U}_{N,[a,b]} = \infty) = 0$. □

(4.31) Proof of Theorem (4.25)

For $a, b \in \mathbb{Q}$, let

$$\Lambda_{a,b} := \{\omega \in \Omega : \lim_{N\to\infty} \mathcal{U}_{N,[a,b]}(\omega) = \infty\}, \qquad \text{and} \qquad \Omega_1 := \Omega \setminus \bigcup_{a,b\in\mathbb{Q}} \Lambda_{a,b}.$$

By (4.30) and σ-additivity, $\mathbb{P}(\Omega_1) = 1$. Since $\mathcal{U}_{N,[a,b]}(\omega) < \infty$ for all $a, b \in \mathbb{Q}$ for all $\omega \in \Omega_1$ by construction, (4.28) implies that $\lim_{n\to\infty} X_n(\omega)$ exists for all $\omega \in \Omega_1$ in $[-\infty, \infty]$. Now,

$$\mathbb{E}(\lim_{n\to\infty} |X_n|) \overset{\text{limit exists a.s.}}{=} \mathbb{E}(\liminf_{n\to\infty} |X_n|) \overset{\text{Fatou}}{\leqslant} \liminf_{n\to\infty} \mathbb{E}(|X_n|) \leqslant \sup_n \mathbb{E}(|X_n|) < \infty.$$

So, $\mathbb{P}(\lim_{n\to\infty} |X_n| < \infty) = 1$. □

(4.32) Corollary

Theorem (4.25) holds for submartingales (X_n).
Proof: Apply it to $(-X_n)$. □

(4.33) Corollary

Let (X_n) be a nonnegative supermartingale. Then $\mathbb{P}(\lim_{n\to\infty} X_n$ exists and is finite$\} = 1$.

Proof: $\mathbb{E}(|X_n|) = \mathbb{E}(X_n) \leqslant \mathbb{E}(X_1)$ since (X_n) is a supermartingale. Thus (X_n) is uniformly bounded in L^1, and Theorem (4.25) applies. □

(4.34) Example

Recall the exponential distribution: $X \sim \text{Exp}(\beta)$ means that $\mathbb{P}(X > \alpha) = \mathrm{e}^{-\alpha\beta}$. Now let (Y_i) be iid and $Y_i \sim \text{Exp}(1)$ for all i. Define recursively

$$X_1 = Y_1, \quad \bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_{n+1}(\omega) = \bar{S}_n(\omega) Y_{n+1}(\omega).$$

Do the following exercises:

a) Check as an exercise that $\mathcal{F}_n := \sigma(X_i : i \leqslant n) = \sigma(Y_i : i \leqslant n)$, and that

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \bar{S}_n \quad \text{a.s.}$$

b) Using $\bar{S}_{n+1} = \frac{n}{n+1}\bar{S}_n + \frac{1}{n+1}X_{n+1}$, conclude that $(\bar{S}_n)$ is a martingale.

c) Show that $\lim_{n\to\infty} \bar{S}_n$ exists a.s.

Note that if the (X_n) would be iid and integrable, then the strong law of large numbers would give that $\bar{S}_n$ converges almost surely, and would also identify the limit, namely the constant RV $\mathbb{E}(X_1)$. In the present case, the limit will not be constant (why?). Can you find out anything about it or its distribution?

Another very useful property of martingales is that the distribution of the maximum up to n is governed by the distribution of the $n-th$ step.

(4.35) Doobs Submartingale Inequality

Let (X_n) be a submartingale. Then for all $c > 0$, $n \in \mathbb{N}$,

$$\mathbb{P}(\max_{k \leqslant n} X_k \geqslant c) \leqslant \frac{1}{c}\mathbb{E}(X_n 1_{\{\max_{k \leqslant n} X_k \geqslant c\}}).$$

In particular, if (X_n) is nonnegative, then $\mathbb{P}(\max_{k \leqslant n} X_k \geqslant c) \leqslant \frac{1}{c}\mathbb{E}(X_n)$.

Proof: For $k \in \mathbb{N}_0$, put

$$A_k = \{X_k \geqslant c, X_j < c \,\forall j < k\}.$$

Then $A_k \in \mathcal{F}_k$. We have for $n \geqslant k$:

$$\mathbb{E}(X_n 1_{A_k}) = \mathbb{E}(\mathbb{E}(X_n 1_{A_k}|\mathcal{F}_k)) = \mathbb{E}(1_{A_k}\mathbb{E}(X_n|\mathcal{F}_k)) \overset{\text{submartingale}}{\geqslant} \mathbb{E}(1_{A_k}X_k) \overset{X_k 1_{A_k} \geqslant c 1_{A_k}}{\geqslant} c\mathbb{P}(A_k).$$

Since $\{\max_{k \leqslant n} X_k \geqslant c\}$ is the disjoint union of the A_k with $k \leqslant n$, summing the above inequality over $k \leqslant n$ gives the result. □

The next lemma is simple but extremely useful. The short version is that convex functions of martingales are submartingales. The long version is

(4.36) Lemma

Let (X_n) be a martingale, $\varphi : \mathbb{R} \to \mathbb{R}$ a convex function with $\mathbb{E}(|\varphi(X_n)|) < \infty$ for all n. Define $Y_n := \varphi(X_n)$. Then (Y_n) is a submartingale.

Proof:

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \overset{\text{Jensen}}{\geqslant} \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \overset{\text{martingale}}{=} \varphi(X_n) = Y_n.$$

□

Remark: In (4.36) the assumption $\mathbb{E}(|\varphi(X_n)|) < \infty$ for all n may be weakened by the condition $\mathbb{E}(\varphi(X_n))^+ < \infty$ for all n. (Exercise)

(4.37) Corollary

Let (X_n) be a martingale. Then for all $c > 0$, $n \in \mathbb{N}$,

$$\mathbb{P}(\max_{k \leqslant n} |X_k| \geqslant c) \leqslant \frac{1}{c}\mathbb{E}(|X_n| 1_{\{\max_{k \leqslant n} |X_k| \geqslant c\}}) \leqslant \frac{1}{c}\mathbb{E}(|X_n|).$$

Proof: $x \mapsto |x|$ is convex. Use (4.36) and (4.35). □

(4.38) Reminder and Remark

For a RV X, its L^p-norm is defined as

$$\|X\|_p := \mathbb{E}(|X|^p)^{1/p} \leqslant \infty.$$

You may know from measure theory that this is a norm on the set of (equivalence classes of) functions for which it is finite.

(4.39) Doobs L^p-maximal inequality

Let (X_n) be a nonnegative submartingale. Then for all $n \in \mathbb{N}$, for all $p > 1$:

$$\| \max_{k \leqslant n} X_k \|_p \leqslant \frac{p}{p-1} \|X_n\|_p.$$

Proof: For nonnegative RVs Z and Y, $r > 0$, we have (exercise!) that

$$(*) \qquad \mathbb{E}(ZY^r) = r \int_0^\infty t^{r-1} \mathbb{E}(Z 1_{\{Y \geqslant t\}}) \, \mathrm{d}t.$$

Thus,

$$\begin{aligned}
\mathbb{E}(|\max_{k \leqslant n} X_k|^p) &\overset{(*),Z=1}{=} p \int_0^\infty t^{p-1} \mathbb{P}(\max_{k \leqslant n} X_k \geqslant t) \, \mathrm{d}t \\
&\overset{(4.35)}{\leqslant} p \int_0^\infty t^{p-1} \tfrac{1}{t} \mathbb{E}(X_n 1_{\{\max_{k \leqslant n} X_k \geqslant t\}}) \, \mathrm{d}t \overset{(*),Z=X_n}{=} \tfrac{p}{p-1} \mathbb{E}(X_n (\max_{k \leqslant n} X_k)^{p-1}) \\
&\overset{\text{Hölder},q=p/(p-1)}{\leqslant} \|X_n\|_p \|(\max_{k \leqslant n} X_k)^{p-1}\|_q = \tfrac{p}{p-1} \|X_n\|_p \mathbb{E}((\max_{k \leqslant n} X_k)^{(p-1)q})^{1/q}.
\end{aligned}$$

Assume first that $\mathbb{E}((\max_{k \leqslant n} X_k)^p) < \infty$. Then using that $(p-1)q = p$ and $1/q = 1 - 1/p$, we divide both sides by $\mathbb{E}((\max_{k \leqslant n} X_k)^p)^{1-1/p}$ to obtain the claim. For the general case, fix $K < \infty$ and let $T = \inf\{n \in \mathbb{N} : X_n \geqslant K\}$. Then T is a stopping time, and thus $(X_{T \wedge n})$ is a submartingale. We thus have

$$\mathbb{E}(|\max_{k \leqslant n} X_{T \wedge k}|^p) \leqslant \tfrac{p}{p-1} \, \|X_n\|_p \, \mathbb{E}(\max_{k \leqslant n} X_{T \wedge k}^p)^{1/q} \tag{4.1}$$

by the calculation above, and

$$\begin{aligned}
\mathbb{E}(\max_{k \leqslant n} X_{T \wedge k}^p) &= \mathbb{E}(1_{\{\max_{k \leqslant n} X_{T \wedge k} \leqslant K\}} \cdot \max_{k \leqslant n} X_{T \wedge k}^p) + \mathbb{E}(1_{\{\max_{k \leqslant n} X_{T \wedge k} > K\}} \cdot \max_{k \leqslant n} X_{T \wedge k}^p) \\
&\leqslant K^p + \mathbb{E}(X_{T \wedge n}^p) \qquad (4.2)
\end{aligned}$$

by definition of T. Now we distinguish two cases. In the first case, let $\mathbb{E}(X_{T \wedge n}^p) < \infty$. Then (4.2) implies $\mathbb{E}(\max_{k \leqslant n} X_{T \wedge k}^p) < \infty$ and dividing both sides of (4.1) by $\mathbb{E}(\max_{k \leqslant n} X_{T \wedge k}^p)^{1-1/p}$ yields

$$(\mathbb{E}(|\max_{k \leqslant n} X_{T \wedge k}|^p))^{1/p} \leqslant \tfrac{p}{p-1} \, \|X_n\|_p.$$

To finish in this case the proof, let $K \to \infty$ and use monotone convergence. For the second case assume $\mathbb{E}(X^p_{T\wedge n}) = \infty$. Using monotone convergence and the conditional Jensen inequality, we obtain

$$\mathbb{E}(X_n^p) = \mathbb{E}(\mathbb{E}(X_n^p|\mathcal{F}_{n-1})) = \lim_{M\to\infty} \mathbb{E}(\mathbb{E}((X_n \wedge M)^p|\mathcal{F}_{n-1}))$$
$$\geqslant \lim_{M\to\infty} \mathbb{E}(\mathbb{E}(X_n \wedge M|\mathcal{F}_{n-1})^p) = \mathbb{E}(\mathbb{E}(X_n|\mathcal{F}_{n-1})^p) \geqslant \mathbb{E}(X_{n-1})^p.$$

Iteratively, this gives: $\mathbb{E}(X_n^p) \geqslant \mathbb{E}(X_m^p)$ for all $m \leqslant n$. Thus, in the second case it follows $\mathbb{E}(X_n^p) = \infty$. This means $||X_n||_p = \infty$, in which case our claim is trivial. □

We end this section by looking at classes of martingales with extra integrability properties.

(4.40) Definition

A martingale (M_n) is called *bounded in* L^2 (or: *an* L^2*-martingale*) if $\sup\{\|M_n\|_2 : n \in \mathbb{N}\} < \infty$.

(4.41) Proposition

Let (M_n) be a martingale with $M_n \in L^2$ for all n. Then

a) the increments of (M_n) are orthogonal i.e.,

$$\forall i \leqslant j \leqslant k \leqslant l : \qquad \langle M_j - M_i, M_l - M_k\rangle := \mathbb{E}((M_j - M_i)(M_l - M_k)) = 0.$$

b) For all n,

$$\mathbb{E}(M_n^2) = \mathbb{E}(M_0^2) + \sum_{m=1}^{n} \mathbb{E}((M_m - M_{m-1})^2)$$

Proof: We have

$$\mathbb{E}((M_j - M_i)(M_l - M_k)) = \mathbb{E}((M_j - M_i)\mathbb{E}(M_l - M_k|\mathcal{F}_k)) = \mathbb{E}((M_j - M_i)\underbrace{(\mathbb{E}(M_l|\mathcal{F}_k) - M_k)}_{=0}) = 0,$$

so a) holds. For b), we have

$$\mathbb{E}(M_n^2) = \mathbb{E}\Big(\big(M_0 + \sum_{m=1}^{n}(M_m - M_{m-1})\big)^2\Big) = \mathbb{E}(M_0^2) + 2\mathbb{E}(M_0\sum_{m=1}^{n}(M_m - M_{m-1}))$$
$$+ \sum_{m=1}^{n}\sum_{p=1}^{m} \mathbb{E}\big((M_m - M_{m-1})(M_p - M_{p-1})\big).$$

By a), the mixed terms in the last line vanish, and the last term in the first line is 0. □

(4.42) Theorem

Let (M_n) be a martingale with $M_n \in L^2$ for all n. Then

a) (M_n) is an L^2-martingale if and only if $\sum_{n=1}^{\infty} \mathbb{E}\big((M_n - M_{n-1})^2\big) < \infty$.

b) For any L^2-martingale, $M := \lim_{n\to\infty} M_n$ exists a.s. and in L^2.

Proof: a) follows directly from (4.41). For b), note that $\mathbb{E}(|M_n|) \leqslant \mathbb{E}(|M_n|^2) + 1$, and since (M_n) is bounded in L^2 we have that $\sup_n \mathbb{E}(|M_n|) < \infty$. Thus almost sure convergence follows from Theorem (4.25).

For L^2-convergence, let M_∞ be the almost sure limit. Then by Fatous Lemma,

$$\mathbb{E}((M_\infty - M_n)^2) = \mathbb{E}(\liminf_{k\to\infty}(M_{n+k} - M_n)^2) \leqslant \liminf_{k\to\infty} \mathbb{E}((M_{n+k} - M_n)^2) = (*).$$

Applying Proposition (4.41), we find

$$(*) = \liminf_{k\to\infty} \sum_{j=n+1}^{n+k} \mathbb{E}((M_j - M_{j-1})^2) = \sum_{j=n+1}^{\infty} \mathbb{E}((M_j - M_{j-1})^2) \xrightarrow{n\to\infty} 0.$$

The claim is shown. □

L^2-martingales can be used to discover interesting facts about sums of independent RVs:

(4.43) Lemma

Let (X_n) be indep. RVs with $\mathbb{E}(X_n) = 0$ and $\sigma_n^2 := \mathbb{V}(X_n) < \infty$ for all n. Define

$$M_n := \sum_{i=1}^n X_i, \qquad A_n := \sum_{i=1}^n \sigma_i^2.$$

Then (M_n) and $(M_n^2 - A_n)$ are martingales.

Proof: exercise. □

(4.44) Theorem

Let (X_n) be indep. RVs with $\mathbb{E}(X_n) = 0$ and $\sigma_n^2 := \mathbb{V}(X_n) < \infty$ for all n.

a) If $\sum_{n=1}^\infty \sigma_n^2 < \infty$, then $\sum_{n=1}^\infty X_n$ exists a.s.

b) Assume that the X_n are a.s. uniformly bounded, i.e. that there exists $K < \infty$ with $X_n(\omega) < K$ for all n and almost all ω. Then the reverse implication to a) also holds, i.e. in this case

$$\sum_{n=1}^\infty \sigma_n^2 < \infty \quad \Leftrightarrow \quad \sum_{n=1}^\infty X_n \text{ exists a.s.}$$

Proof: a) Set $M_n := \sum_{i=1}^n X_i$. (M_n) is an L^2-martingale, since $\mathbb{E}((M_{n+1} - M_n)^2) = \mathbb{E}(X_{n+1}^2) = \sigma_{n+1}^2$, now use (4.42 a). So claim a) follows.

For b), only $\Leftarrow$ remains to be shown. Define (M_n) and (A_n) as in (4.43). The idea of the proof is as follows: we assumed that (M_n) converges a.s. Now $(M_n^2 - A_n)$ is also a martingale, and uniform boundedness of the X_i means that it has a very good chance to converge, too. Then (A_n) also converges. The implementation of 'has a very good chance' needs a bit of care.

Fix $c < \infty$, and define

$$T(\omega) \equiv T_c(\omega) := \inf\{k \in \mathbb{N} : |M_k(\omega)| > c\}.$$

T is an $(\mathcal{F}_n)$-stopping time, and $N_n := M_n^2 - A_n$ is an $(\mathcal{F}_n)$-martingale. Thus by Theorem (4.18), $(N_{T\wedge n})$ is a martingale, and thus

$$\mathbb{E}(M_{T\wedge n}^2) = \mathbb{E}(\sum_{j=1}^{T\wedge n} \sigma_j^2).$$

By the definition of T, $M_{(T\wedge n)-1} \leqslant c$ for all n. By the uniform boundedness of the X_j,

$$M_{T(\omega)\wedge n}(\omega) = M_{(T(\omega)\wedge n)-1}(\omega) + X_{T(\omega)\wedge n}(\omega) \leqslant c + K$$

almost surely, and thus

$$\mathbb{E}(\sum_{j=1}^{T\wedge n} \sigma_j^2) = \mathbb{E}(M_{T\wedge n}^2) \leqslant (c+K)^2$$

for all n. Since (M_n) converges a.s. (by assumption!), we find that

$$1 = \mathbb{P}(\{\omega : (M_n(\omega))_{n\in\mathbb{N}} \text{ is bounded}) = \mathbb{P}(\exists c > 0 : \forall N \in \mathbb{N} : |M_N| < c) =$$
$$= \mathbb{P}(\exists c > 0 : T_c = \infty) \overset{\text{Monotonicity}}{=} \mathbb{P}(\bigcup_{c=1}^{\infty} \{\omega \in \Omega : T_c(\omega) = \infty\}).$$

Thus, there exists some $c \in \mathbb{N}$ with $\mathbb{P}(T_c = \infty) = p_0 > 0$. We conclude that for this c and all $n \in \mathbb{N}$,

$$\sum_{i=1}^{n} \sigma_i^2 = \frac{1}{p_0}\mathbb{E}(\sum_{i=1}^{n} \sigma_i^2 1_{\{T_c=\infty\}}) = \frac{1}{p_0}\mathbb{E}(\sum_{i=1}^{T_c\wedge n} \sigma_i^2 1_{\{T_c=\infty\}}) \leqslant \frac{1}{p_0}\mathbb{E}(\sum_{i=1}^{T_c\wedge n} \sigma_i^2) \leqslant \frac{(c+K)^2}{p_0}.$$

The proof is finished. □

(4.45) Corollary

Let (X_n) be a.s. uniformly bounded, independent RVs. Then

$$\mathbb{P}(\sup_{n\in\mathbb{N}} |\sum_{i=1}^{n} X_i| < \infty) > 0 \quad \Leftrightarrow \quad \mathbb{P}(\lim_{n\to\infty} \sum_{i=1}^{n} X_i \text{ exists in } \mathbb{R}) = 1.$$

Proof: follows from the proof of Theorem (4.44); exercise. □

(4.46) Example: Random signs

Theorem (4.44) has the following nice application. You know from Analysis 1 that there are sequences (a_n) with $a_n \geqslant 0$ for all n, where $\lim_{n\to\infty} a_n = 0$ but $\sum_{n=1}^{\infty} a_n = \infty$. In the class of sequences where $a_n = n^{-\gamma}$, this happens if and only if $\gamma \leqslant 1$.

On the other hand, if the sequence (a_n) is also monotone, then the alternating version $\sum_{i=1}^{\infty}(-1)^n a_n$ always converges, no matter how slowly (a_n) goes to zero (Leibnitz theorem). Of course, the reason is that positive and negative contributions cancel because of the extremely regular pattern of positive and negative signs. So when $a_n = n^{-\gamma}$, the alternating series converges for all $\gamma > 0$.

An interesting question is what happens with random signs: Let (ε_n) be iid with $\mathbb{P}(\varepsilon_n = \pm 1) = 1/2$. What is the right condition for $\sum_{n=1}^\infty \varepsilon_n a_n$ to converge (a.s.)? The answer is given by Theorem (4.44): we set $X_n := \varepsilon_n a_n$ and obtain

$$\sum_{n=1}^\infty \varepsilon_n a_n \text{ exists a.s.} \qquad \Leftrightarrow \qquad \sum_{n=1}^\infty a_n^2 < \infty.$$

Thus, sequences of the type $a_n = n^{-\gamma}$ with random signs converge if and only if $\gamma > 1/2$.

The proof of the following statement about random signs is left as an exercise:

$$\limsup_{n\to\infty} \sum_{i=1}^n \varepsilon_i a_i = -\liminf_{n\to\infty} \sum_{i=1}^n \varepsilon_i a_i = \infty \text{ a. s.} \qquad \Leftrightarrow \qquad \sum_{n=1}^\infty a_n^2 = \infty.$$

(4.47) Remark

We know that when (M_n) is a martingale with $M_n \in L^2$ for all n, then (M_n^2) is a submartingale (why?). So (M_n^2) is 'increasing on average': $\mathbb{E}(M_n^2|\mathcal{F}_{n-1}) \geqslant M_{n-1}^2$ almost surely. Lemma (4.43) says that when (M_n) is the sum of n independent square integrable RVs, then we can correct (M_n^2) back to a martingale by subtracting a strictly increasing process, namely the deterministic process (A_n) which is the sum of the first n variances. The next result is not hard to prove, but it is remarkable because it says that for a completely arbitrary L^2-martingale, such a monotone increasing correction of the submartingale (M_n^2) to a martingale is possible, it is unique, and we can give a formula. We first give the axiomatic definition.

(4.48) Definition

Let (M_n) be a martingale with respect to the σ-algebra $(\mathcal{F}_n)$ with $M_n \in L^2$ for all n. Any process with the properties

(i): (A_n) is $(\mathcal{F}_n)-$previsible, i.e. A_n is $\mathcal{F}_{n-1}$-measurable for all n,

(ii): $A_0 = 0$, and (A_n) is increasing in n a.s., i.e. the sequence $n \mapsto A_n(\omega)$ is monotone increasing for almost all ω,

(iii): $(M_n^2 - A_n)$ is an $(\mathcal{F}_n)$-martingale,
is called a *quadratic variation* of the martingale (M_n).

(4.49) Theorem

In the situation of (4.48), there exists a unique (a.s.) process that has properties (i)-(iii). This process is called *the* quadratic variation process for (M_n). It is given by the formula

$$A_n = \sum_{k=1}^n \mathbb{E}((M_k - M_{k-1})^2|\mathcal{F}_{k-1}) = \sum_{k=1}^n \mathbb{E}(M_k^2 - M_{k-1}^2|\mathcal{F}_{k-1}).$$

Proof: We first show that the two expressions in the formula coincide:

$$\mathbb{E}((M_k - M_{k-1})^2|\mathcal{F}_{k-1}) = \mathbb{E}(M_k^2|\mathcal{F}_{k-1}) + M_{k-1}^2 - 2M_{k-1}\underbrace{\mathbb{E}(M_k|\mathcal{F}_{k-1})}_{=M_{k-1}} = \mathbb{E}(M_k^2 - M_{k-1}^2|\mathcal{F}_{k-1}).$$

Now we show that (A_n) is a quadratic variation. From the first expression on the right hand side of the formula we read off that $A_0 = 0$, and that A_n is increasing and previsible. From the second expression we conclude that

$$\mathbb{E}(M_n^2|\mathcal{F}_{n-1}) - M_{n-1}^2 = \mathbb{E}(M_n^2 - M_{n-1}^2|\mathcal{F}_{n-1}) = A_n - A_{n-1}.$$

Using that (A_n) is previsible and rearranging gives

$$\mathbb{E}(M_n^2 - A_n|\mathcal{F}_{n-1}) = M_{n-1}^2 - A_{n-1},$$

which shows that (A_n) is a quadratic variation.

Finally we show uniqueness. Let $(\tilde{A}_n)$ be any process fulfilling (i)-(iii). Then for all $1 \leqslant k \leqslant n$,

$$\mathbb{E}(M_k^2 - \tilde{A}_k|\mathcal{F}_{k-1}) = M_{k-1}^2 - \tilde{A}_{k-1},$$

and therefore (by previsibility)

$$\tilde{A}_k - \tilde{A}_{k-1} = \mathbb{E}(M_k^2 - M_{k-1}^2|\mathcal{F}_{k-1}).$$

Summing this up from $k = 1$ to n, and using $\tilde{A}_0 = 0$, we find that $\tilde{A}_n = A_n$ almost surely. □

(4.50) Notation

If (M_n) is a martingale with $M_n \in L^2$ for all n, we often write $(\langle M_n \rangle)_{n\in\mathbb{N}}$ for its quadratic variation.

If we give up the monotonicity requirement, we have an even more general decomposition into a martingale and a previsible process:

(4.51) Doob Decomposition

Let (X_n) be an $(\mathcal{F}_n)$-adapted process with $X_n \in L^1$ for all n. Then there exists an a.s. unique pair $((M_n), (A_n))$ where (M_n) is a martingale with $M_0 = 0$, (A_n) is a previsible process with $A_0 = 0$, and for all n,

$$X_n = X_0 + M_n + A_n \qquad \text{almost surely.}$$

Here, (A_n) and (M_n) are given by the formulae

$$A_n = \sum_{k=1}^{n} \mathbb{E}(X_k - X_{k-1}|\mathcal{F}_{k-1}), \qquad M_n = X_n - X_0 - A_n.$$

Proof: Imitate the proof of Theorem (4.49). Exercise. □

We will now see one reason why the quadratic variation is very useful. (another one will come in the module Stochastic Processes where we do stochastic integrals).

(4.52) Lemma

Let (M_n) be a martingale with $M_n \in L^2$ for all n and $M_0 = 0$. Then $\langle M_\infty\rangle(\omega) := \lim_{n\to\infty}\langle M_n\rangle(\omega)$ exists a.s. in $[0, \infty]$, and

$$\mathbb{E}(\langle M_\infty\rangle) < \infty \quad \Leftrightarrow \quad (M_n) \text{ is an } L^2\text{-martingale.}$$

Proof: Since $n \mapsto \langle M_n\rangle(\omega)$ is increasing a.s., the limit $\langle M_\infty\rangle(\omega)$ exists a.s. Monotone convergence gives

$$\mathbb{E}(\langle M_\infty\rangle) = \lim_{n\to\infty} \mathbb{E}(\langle M_n\rangle) = \sup_{n\in\mathbb{N}} \mathbb{E}(\langle M_n\rangle) = \sup_{n\in\mathbb{N}} \mathbb{E}(M_n^2).$$

The last equality holds (even without the $\sup_n$) because $(M_n^2 - \langle M_n\rangle)$ is a martingale. The claim is shown. □

The following Theorem should be compared with the Martingale convergence Theorem (4.25). Note in particular that uniform L^1-boundedness is not assumed below.

(4.53) Theorem

Let (M_n) be a martingale with $M_0 = 0$ and $M_n \in L^2$ for all n. Write $A_n = \langle M_n\rangle$, and $A_\infty = \lim_{n\to\infty} A_n \in [0, \infty]$.

a) $\lim_{n\to\infty} M_n$ exists a.s. on the set $\{A_\infty < \infty\}$, i.e.

$$\mathbb{P}(A_\infty < \infty, \lim_{n\to\infty} M_n \text{ does not exist}) = 0.$$

b) If we assume in addition that (M_n) has a.s. uniformly bounded increments (i.e. that for some $K < \infty$, we have $|M_{n+1} - M_n| \leqslant K$ a.s.), then the following converse of a) is also true: $A_\infty < \infty$ a.s. on the set $\{\lim_{n\to\infty} M_n \text{ exists.}\}$. In other words, in this case

$$\mathbb{P}(A_\infty = \infty, \lim_{n\to\infty} M_n \text{ exists}) = 0.$$

c) A reformulation of b): In the situation of b), we can find a set Ω_0 of measure zero so that for all $\omega \notin \Omega_0$, $\lim_{n\to\infty} M_n(\omega)$ exists if and only if $A_\infty(\omega) < \infty$. Note: We make no statement about the value of $\mathbb{P}(\lim_{n\to\infty} M_n \text{ exists})$.

Proof: a) For $k \in \mathbb{N}$, we define

$$S(k, \omega) := \inf\{n \in \mathbb{N}_0 : A_{n+1}(\omega) > k\}.$$

For each k, the map $\omega \mapsto S(k, \omega)$ is a stopping time, since

$$\{S(k, .) > n\} = \{\forall j \leqslant n : A_{j+1} \leqslant k\} = \bigcap_{j=0}^{n} A_{j+1}^{-1}((-\infty, k]) \in \mathcal{F}_n.$$

In the last statement, we used that (A_n) is previsible. Since $S(k, .)$ is a stopping time, the process $(M_{S(k,.)\wedge n})$ is a martingale. We claim that

$$\langle M_{S(k,.)\wedge n}\rangle = A_{S(k,.)\wedge n}. \qquad (*)$$

To see $(*)$, first note that since $(M_n^2 - A_n)$ is a martingale, the optional stopping time guarantees that $(M_{S(k,.)\wedge n}^2 - A_{S(k,.)\wedge n})$ is a martingale. Since $A_{S(k,.)\wedge 0} = A_0 = 0$ and $n \mapsto A_{S(k,\omega)\wedge n}(\omega)$ is increasing, it remains to show that $(A_{S(k,.)\wedge n})$ is previsible. This is true since (A_n) itself is previsible, and thus

$$\{A_{S(k,.)\wedge n} \leqslant C\} = \bigcup_{m=1}^{n-1} \{S(k,.) = m, A_m \leqslant C\} \cup \big(\{S(k,.) \leqslant n-1\}^c \cap \{A_n \leqslant C\}\big) \in \mathcal{F}_{n-1}.$$

So $(*)$ holds, and

$$\mathbb{E}(\lim_{n\to\infty} A_{S(k,.)\wedge n}) = \mathbb{E}(\underbrace{A_{S(k,.)}}_{\leqslant k} 1_{\{A_{S(k,.)<\infty}\}} + \underbrace{A_\infty 1_{\{A_{S(k,.)=\infty}\}}}_{\leqslant k}) \leqslant 2k < \infty.$$

By Lemma (4.52), we conclude that for all k, the martingale $(M_{S(k,.)\wedge n})$ is bounded in L^2 and thus converges a.s. This means that for each k,

$$\mathbb{P}\big(S(k,.) = \infty, \lim_{n\to\infty} M_n \text{ does not exist}\big) = \mathbb{P}\big(S(k,.) = \infty, \lim_{n\to\infty} M_{S(k,.)\wedge n} \text{ does not exist}\big) = 0.$$

Since

$$\{A_\infty < \infty\} = \{\exists k \in \mathbb{N} : S(k,\omega) < \infty\} = \bigcup_{k=1}^{\infty} \{S(k,.) = \infty\},$$

a) now follows from

$$\mathbb{P}(A_\infty < \infty, \lim_{n\to\infty} M_n \text{ does not exist}) \leqslant \sum_{k=1}^{\infty} \mathbb{P}(S(k,.) = \infty, \lim_{n\to\infty} M_n \text{ does not exist}) = 0.$$

For b), note first that

$$\mathbb{P}(A_\infty = \infty, \lim_{n\to\infty} M_n \text{ exists}) \leqslant \mathbb{P}(A_\infty = \infty, \sup_{n\in\mathbb{N}} |M_n| < \infty) =: p.$$

Assume now that b) is false and thus $p > 0$. Then there exists $c > 0$ so that

$$(**) \quad \mathbb{P}(A_\infty = \infty, T(c,.) = \infty) > 0 \qquad \text{where } T(c,\omega) = \inf\{n \in \mathbb{N} : |M_n(\omega)| > c\}.$$

Since $(M^2_{T(c,.)\wedge n} - A_{T(c,.)\wedge n})$ is a martingale, we have

$$\mathbb{E}(A_{T(c,.)\wedge n}) = \mathbb{E}(M^2_{T(c,.)\wedge n}) \leqslant \mathbb{E}((M_{T(c,.)\wedge n-1} + K)^2) \leqslant (c+K)^2.$$

Taking $n \to \infty$ and using monotone convergence gives $\mathbb{E}(A_{T(c,.)}) < (c+K)^2$, and so

$$1 = \mathbb{P}(A_\infty < \infty) = \mathbb{P}(A_\infty < \infty, T(c,.) = \infty) + \mathbb{P}(A_\infty < \infty, T(c,.) < \infty) =$$
$$= \mathbb{P}(T(c,.) = \infty) - \mathbb{P}(A_\infty = \infty, T(c,.) = \infty) + \mathbb{P}(T(c,.) < \infty) = 1 - \mathbb{P}(A_\infty = \infty, T(c,.) = \infty).$$

This means that $\mathbb{P}(A_\infty = \infty, T(c,.) = \infty) = 0$, in contradiction to $(**)$. So, b) must hold. c) now is a direct consequence. □

As a preparation for the next Theorem, we need the following classical results from real analysis:

(4.54) Cesàros Lemma

Let (v_n) be a convergent sequence with limit v_∞, and (c_n) a nonnegative sequence with $c_1 > 0$ and $\sum_{n=1}^\infty c_n = \infty$. Then

$$\lim_{n\to\infty} \frac{1}{\sum_{i=1}^n c_i} \sum_{k=1}^n c_k v_k = v_\infty.$$

In words: weighted averages of convergent sequences converge to the same limit as the original sequence.

Proof: exercise. □

(4.55) Kroneckers Lemma

For any monotone decreasing sequence (a_n) with $\lim_{n\to\infty} a_n = 0$, and any real sequence (x_n) the following statement holds:

$$\text{If } \lim_{n\to\infty} \sum_{k=1}^n a_k x_k \text{ exists, then } \lim_{n\to\infty} a_n \sum_{k=1}^n x_k = 0.$$

Proof: Put $u_n = \sum_{k=1}^n a_k x_k$. We have

$$\sum_{k=1}^n x_k = \sum_{k=1}^n \frac{1}{a_k}(u_k - u_{k-1}) = \frac{u_n}{a_n} - \sum_{k=1}^n \Big(\frac{1}{a_k} - \frac{1}{a_{k-1}}\Big) u_{k-1}.$$

Thus with $k_n := \frac{1}{a_n}$, $c_n := z_n - z_{n-1}$ we have $z_n = \sum_{k=1}^n c_k$ and

$$a_n \sum_{k=1}^n x_k = u_n - \frac{1}{z_n} \sum_{k=1}^n c_k u_{k-1}.$$

Since we assumed the convergence of (u_n), both terms above converge to $\lim_{n\to\infty} u_n$, the second one by Cesàros Lemma. The claim follows. □

(4.56) Theorem (SLLN for martingales)

Let (M_n) be a martingale, $M_0 = 0$, and $M_n \in L^2$ for all n. Then

$$\lim_{n\to\infty} \frac{1}{\langle M_n\rangle(\omega)} M_n(\omega) = 0 \quad \text{for almost all } \omega \in \{\langle M_\infty\rangle = \infty\}.$$

In particular, $\lim_{n\to\infty} \frac{1}{\langle M_n\rangle(\omega)} M_n(\omega)$ exists almost surely.

Proof: We write $A_n = \langle M_n\rangle$ and define

$$X_n := (\frac{1}{1+A} \bullet M)_n = \sum_{k=1}^n \frac{1}{1+A_k}(M_k - M_{k-1}).$$

Since (A_n) is previsible, so is $(1/(1+A_n))$, and thus (X_n) is a martingale. We estimate $\langle X_n \rangle$:

$$\begin{aligned}\langle X_n \rangle &= \sum_{k=1}^n \mathbb{E}((X_k - X_{k-1})^2|\mathcal{F}_{k-1}) = \sum_{k=1}^n \mathbb{E}\Big(\frac{1}{(1+A_k)^2}(M_k - M_{k-1})^2\Big|\mathcal{F}_{k-1}\Big) = \\ &= \sum_{k=1}^n \frac{1}{(1+A_k)^2} \underbrace{\mathbb{E}((M_k - M_{k-1})^2|\mathcal{F}_{k-1})}_{=A_k - A_{k-1}} = \sum_{k=1}^n \frac{A_k - A_{k-1}}{(1+A_k)^2} \leqslant \sum_{k=1}^n \frac{A_k - A_{k-1}}{(1+A_k)(1+A_{k-1})} \\ &= \sum_{k=1}^n \Big(\frac{1}{1+A_{k-1}} - \frac{1}{1+A_k}\Big) = 1 - \frac{1}{1+A_n}.\end{aligned}$$

This means that $\langle X_\infty \rangle \leqslant 1$ a.s., and so by Theorem (4.53 a), $\lim_{n\to\infty} X_n$ exists a.s.. Now for ω with $A_\infty(\omega) = \infty$, the sequence $a_n := \frac{1}{1+A_n(\omega)}$ converges to zero, and so Kroneckers Lemma (with $x_n = M_n(\omega) - M_{n-1}(\omega)$) implies that

$$\lim_{n\to\infty} \frac{1}{1+A_n(\omega)} M_n(\omega) = 0 \qquad \text{a.s.}$$

The first claim now follows. For the second claim, combine this with Theorem (4.53 a). □

(4.57) Remark

We know from Theorem (4.42 b) that an L^2-martingale converges to its a.s. limit also in L^2. Why is this interesting? Because it tells us that (in this case) the variance of the approximations M_n has anything to do with the variance of the limit M_∞. In the next example, we see that in cases where there is almost sure convergence but not L^1-convergence, unintuitive things happen. Recall that by the martingale convergence theorem, $\sup \|M_n\|_{L^1} < \infty$ implies the existence of $\lim_{n\to\infty} M_n$ almost surely.

(4.58) Example

Consider the following game: a fair coin is thrown n times. If all outcomes are 'head', you get 2^n Euro. If at least one outcome is 'tail', you get nothing. How much would you be willing to pay for the privilege of playing the game? Will your answer depend on n?

There are two answers to this question. Both start with the observation that when (X_n) is a sequence of iid RVs with $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 2) = 1/2$, then the process (M_n) with $M_0 = 1$, $M_n = \prod_{i=1}^n X_i$ is a martingale, see Example (4.9 b).

a) First answer: Since (M_n) is a martingale, with an initial 'wealth' of $M_0 = 1$ the game is fair. So, independent of n, you should pay 1 Euro to play the game. This point of view is supported by the fact that $\mathbb{E}(M_n) = 1$ for all n.

b) Second answer: Your chance of winning anything at all is 2^{-n}. Of course, if you do win, you win an astronomical sum for large n, but this will never happen when n is 100 or more, for example. So, you should pay a bit for playing at very small n, but definitely not play when n is very large.

The first answer seems more mathematically well-founded, but the second one is the one most people will probably choose. In contrast to other situations, the intuition here is not misleading. The reason is that (M_n) is a martingale with $\mathbb{E}(M_n) = 1$, so it converges a.s.. What is the limit? Of course, it is $M_\infty = 0$. So, almost surely you lose in the long run. Also, this shows that $0 = \mathbb{E}(\lim_n M_n) < \lim \mathbb{E}(M_n) = 1$, an example where Fatou's Lemma is a strict inequality.

The mathematical reason for this is that the distribution of the random variables (M_n) 'hides' almost all of its mass on a tiny subset of the original probability space $\Omega = \{0, 2\}^{\mathbb{N}}$: $M_n = 2^n$ on the set $\{(x_m)_{m \in \mathbb{N}} : x_m = 2 \, \forall m \leqslant n\}$ which has probability 2^{-n}. In the remainder of this chapter, we investigate what happens if we do not allow a martingale to 'hide' its mass. We start with the appropriate definition. You should pay attention to the way in which a situation like above is excluded by it.

(4.59) Definition

Let $(X_n)_{n \in I}$ be a family of random variables (the index set I can be uncountable). The family is called *uniformly integrable* (short: UI) if

$$\forall \varepsilon > 0 : \quad \exists K < \infty : \quad \forall n \in I : \quad \mathbb{E}(|X_n| 1_{\{|X_n| > K\}}) < \varepsilon.$$

(4.60) Proposition

A familiy of finitely many integrable RVs is UI.

Proof: Consider fist a single RV X. Assume that X is not uniformly integrable. Then there exists $\varepsilon > 0$ so that for all $y > 0$, $\mathbb{E}(|X| 1_{\{|X| > y\}}) \geqslant \varepsilon$. Consequently,

$$\mathbb{E}(|X|) = \int_0^\infty \mathbb{P}(|X| > y) \, \mathrm{d}y \geqslant \int_1^\infty \frac{1}{y} \mathbb{E}(|X| 1_{\{X > y\}}) \geqslant \varepsilon \int_1^\infty \frac{1}{y} \mathrm{d}y = \infty,$$

so X is not integrable. Thus the claim holds for a single RV. For finitely many RVs, just take the maximum of all the constants K that you find for each RV individually. This maximum fulfils the UI condition. □

(4.61) Proposition

A family $(X_n)_{n \in I}$ of RVs is UI if one of the two following statements is true.

a) There exists $p > 1$ so that (X_i) is bounded in L^p, i.e. such that $\sup_{n \in I} \mathbb{E}(|X_n|^p) < \infty$.

b) (X_n) is dominated by one integrable RV X, i.e. there exists $X \in L^1$ so that for all $n \in I$, $X_n \leqslant X$ a.s.

Proof: a) Let $\varepsilon > 0$, and choose K with $K^{1-p} \sup_{n \in I} \mathbb{E}(|X|^p) < \varepsilon$. Then since $p > 1$,

$$\mathbb{E}(|X_n| 1_{\{|X_n| > K\}}) = \mathbb{E}(|X_n|^p |X_n|^{1-p} 1_{\{|X_n|^{1-p} < K^{1-p}\}}) \leqslant K^{1-p} \mathbb{E}(|X_n|^p) < \varepsilon.$$

b) We have

$$\sup_{n\in I} \mathbb{E}(|X_n| 1_{\{|X_n|>K\}}) \leqslant \mathbb{E}(|X| 1_{\{|X|>K\}})$$

for all K. The claim now follows from Proposition (4.60). □

(4.62) Example

We will now give an example of a UI martingale. As we will see below, it is essentially the only example. Let $X \in L^1$ be a random variable, and $(\mathcal{F}_n)$ a filtration. Example (4.9 c) shows that (M_n) with $M_n = \mathbb{E}(X|\mathcal{F}_n)$ is a martingale. The interpretation of M_n is that by refining the σ-algebra from $\mathcal{F}_n$ to $\mathcal{F}_{n+1}$, we discover more and more properties of X. Theorem (4.64) below will imply that (M_n) is UI. Before we state it, we need a strengthening of Proposition (4.60).

(4.63) Lemma

Let $X \in L^1$. Then for each $\varepsilon > 0$ we can find $\delta > 0$ so that

$$\sup\{\mathbb{E}(|X| 1_A) : A \in \mathcal{F}, \mathbb{P}(A) < \delta\} < \varepsilon.$$

Proof: Assume the contrary, i.e. that there exists $\varepsilon > 0$ so that for all $\delta > 0$, we can find a set $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$ but $\mathbb{E}(|X| 1_A) > \varepsilon$. Then we pick such a set A_n for each $\delta_n = 2^{-n}$, and define $\bar{A} := \limsup_{n\to\infty} A_n$. Since $\sum_n \mathbb{P}(A_n) < \infty$, the Borel-Cantelli Lemma implies that $\mathbb{P}(\bar{A}) = 0$. This means that

$$\begin{aligned} 0 = \mathbb{E}(|X| 1_{\bar{A}}) = \mathbb{E}(|X| \limsup_{n\to\infty} 1_{A_n}) = \mathbb{E}(|X|) - \mathbb{E}(|X| \liminf_{n\to\infty} 1_{A_n^c}) \overset{\text{Fatou}}{\geqslant} \\ \geqslant \mathbb{E}(|X|) - \liminf_{n\to\infty} \mathbb{E}(|X| 1_{A_n^c}) = \limsup_{n\to\infty} \mathbb{E}(|X| 1_{A_n}) \geqslant \varepsilon. \end{aligned}$$

This contradiction shows the claim. □

(4.64) Theorem

Let X an integrable RV on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$\Sigma := \{\mathcal{G} \subset \mathcal{F} : \mathcal{G} \text{ is a } \sigma\text{-algebra } \}.$$

Then the family $(\mathbb{E}(X|\mathcal{G}))_{\mathcal{G}\in\Sigma}$ is UI.

Proof: Use Jensens inequality first: $|\mathbb{E}(X|\mathcal{G})| \leqslant \mathbb{E}\big(|X| \,\big|\, \mathcal{G}\big)$. Thus

$$\mathbb{E}\Big(\big|\mathbb{E}(X|\mathcal{G})\big| \, 1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}\Big) \leqslant \mathbb{E}\Big(\mathbb{E}\big(|X| \,\big|\, \mathcal{G}\big) 1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}\Big) = \mathbb{E}\big(|X| 1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}\big).$$

In the last equality we used the definition of conditional expectation and the fact that $1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}$ is $\mathcal{G}$-measurable. On the other hand, for all $K > 0$, Chebyshevs inequality gives

$$\mathbb{P}(|\mathbb{E}(X|\mathcal{G})| > K) \leqslant \frac{1}{K}\mathbb{E}(|\mathbb{E}(X|\mathcal{G})|) \leqslant \frac{1}{K}\mathbb{E}(\mathbb{E}(|X| \,|\, \mathcal{G})) = \frac{1}{K}\mathbb{E}(|X|).$$

Now let $\varepsilon > 0$ and choose $\delta > 0$ so that the statement of Lemma (4.63) holds. Picking $K > \mathbb{E}(|X|)/\delta$, we find

$$\sup\Big\{\mathbb{E}\Big(\big|\mathbb{E}(X|\mathcal{G})\big|\,1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}\Big) : \mathcal{G} \in \Sigma\Big\} \leqslant \sup\Big\{\mathbb{E}\big(|X|1_{\{|\mathbb{E}(X|\mathcal{G})|>K\}}\big) : \mathcal{G} \in \Sigma\Big\}$$
$$\leqslant \sup\{\mathbb{E}(X1_A) : \mathbb{P}(A) < \delta\} < \varepsilon.$$

□

We now need two statements that compare different types of convergence. You should compare them to Theorem (1.28).

(4.65) Lemma

Assume that a sequence (X_n) of RVs is uniformly bounded and converges to some RV X in probability. Then $X \in L^1$, and $\lim \|X_n - X\|_1 = 0$.

Proof: Let $\varepsilon > 0$. We have

$$\|X_n - X\|_1 = \mathbb{E}(|X_n - X|) \leqslant \mathbb{E}(|X_n - X|1_{\{|X_n - X|>\varepsilon\}}) + \varepsilon = (*).$$

By assumption, there exists $K < \infty$ with $|X_n| < K$ a.s. for all n. Then for all $m > 1$, $\mathbb{P}(|X| > K + 1/m) \leqslant \mathbb{P}(|X - X_n| > 1/m) \to 0$ as $n \to \infty$. So, $\mathbb{P}(|X| > K + 1/m) = 0$ for all m, and thus also

$$\mathbb{P}(|X| > K) = \mathbb{P}(\bigcup_{m\in\mathbb{N}}\{|X| > K + 1/m\}) = 0.$$

In particular, $X \in L^1$, and

$$(*) \leqslant 2K\mathbb{P}(|X_n - X| > \varepsilon) + \varepsilon \overset{n\to\infty}{\longrightarrow} 0 + \varepsilon.$$

As this is true for arbitrary $\varepsilon > 0$, the claim holds. □

(4.66) Theorem

Let (X_n) be a sequence of integrable RVs, and let X be a RV. Then the following statements are equivalent:

(i): $X \in L^1$ and $\lim_{n\to\infty} \|X_n - X\|_1 = 0$.

(ii): $X_n \to X$ in probability, and (X_n) is UI.

Proof: *(ii)* ⇒ *(i)*: For $K > 0$, define the K-cutoff RV $X_n^{(K)} = \max\{\min\{X_n, K\}, -K\}$. The triangle inequality gives

$$\mathbb{E}(|X_n - X_m|) \leqslant \mathbb{E}(|X_n - X_n^{(K)}|) + \mathbb{E}(|X_n^{(K)} - X_m^{(K)}|) + \mathbb{E}(|X_n - X_m^{(K)}|) \qquad (*)$$

for all n, m. For each K, $(X_n^{(K)})$ converges in probability since (X_n) does, and is bounded. So the middle term in $(*)$ is an L^1-Cauchy sequence by (4.64). Since (X_n) is UI, for each $\varepsilon > 0$ there exists some $K < \infty$ so that

$$\sup\{n \in \mathbb{N} : \mathbb{E}(|X_n^{(K)} - X_n|) = \sup\{n \in \mathbb{N} : \mathbb{E}(|X_n|1_{\{|X_n|>K\}})\} < \varepsilon.$$

Inserting these two facts into $(*)$, we find that for each $\varepsilon > 0$ we can find $N > 0$ so that for all $n \geqslant N$, $\mathbb{E}(|X_n - X_m|) < 3\varepsilon$. So, (X_n) is a L^1-Cauchy sequence and thus converges. Since convergence in L^1 implies convergence in probability (to the same function), the limit must be X.

$(i) \Rightarrow (ii)$: By (1.28 d), we only need to show that (X_n) is UI. For this, let $K < \infty$. Then for all $n \in \mathbb{N}$,

$$\mathbb{E}(|X_n| 1_{\{|X_n|>K\}}) \leqslant \mathbb{E}((|X_n - X| + |X|) 1_{\{|X_n|>K\}}) = \mathbb{E}(|X_n - X| 1_{\{|X_n|>K\}}) + \mathbb{E}(|X| 1_{\{|X_n|>K\}}).$$

Now let $\varepsilon > 0$ and choose $N \in \mathbb{N}$ with $\mathbb{E}(|X_n - X|) < \varepsilon$ for $n > N$. Then

$$\begin{aligned} \sup\{\mathbb{E}(|X_n| 1_{\{|X_n|>K\}}) : m \in \mathbb{N}\} &\leqslant \sup\{\mathbb{E}(|X_n - X| 1_{\{|X_n|>K\}}) : n \leqslant N\} + \\ &+ \sup\{\mathbb{E}(|X_n - X|) : n > N\} + \sup\{\mathbb{E}(|X| 1_{\{|X_n|>K\}}) : n \in \mathbb{N}\} \\ &\leqslant \sup\{\mathbb{E}(|X_n| 1_{\{|X_n|>K\}}) : n \leqslant N\} + \varepsilon + 2 \sup\{\mathbb{E}(|X| 1_{\{|X_n|>K\}}) : n \in \mathbb{N}\}. \qquad (*) \end{aligned}$$

The first term of $(*)$ can be made less than ε by choosing K large enough, since finitely many integrable RVs are UI. For the third term, note that $\mathbb{E}(|X_n|) \leqslant \mathbb{E}(|X_n - X|) + \mathbb{E}(|X|) \to \mathbb{E}(|X|)$, and so we have $\sup_n \mathbb{E}(|X|) < \infty$. By Chebyshev, for each $\delta > 0$ we can choose K so large that

$$\sup\{\mathbb{P}(|X_n| > K) : n \in \mathbb{N}\} \leqslant \sup\{\frac{1}{K}\mathbb{E}(|X_n|) : n \in \mathbb{N}\} < \delta.$$

Now fix $\varepsilon > 0$ and for each n let δ_n and K_n be so small that the statement of Lemma (4.63) holds for $|X|$. Then the third term in $(*)$ is also $\leqslant \varepsilon$, and we have shown that (X_n) is UI. □

We can now apply this to martingales:

(4.67) Theorem

Let (M_n) be a submartingale with its natural filtration, i.e. with $\mathcal{F}_n = \sigma(M_k : k \leqslant n)$. If (M_n) is UI, then $M_\infty = \lim_{n\to\infty} M_n$ exists a.s. and in L^1, and

$$\forall n : \qquad M_n \leqslant \mathbb{E}(M_\infty | \mathcal{F}_n).$$

If (M_n) is a martingale, then even

$$\forall n : \qquad M_n = \mathbb{E}(M_\infty | \mathcal{F}_n).$$

Proof: Since (M_n) is UI, we have $\sup_n \mathbb{E}(|M_n|) < \infty$, and so $M_\infty = \lim_{n\to\infty} M_n$ exists a.s. by the Martingale Convergence Theorem. Thus $M_n \to M_\infty$ in probability, and so $M_\infty \in L^1$ and $M_n \to M_\infty$ in L^1 by Theorem (4.66). For $k \geqslant n$, we compute

$$\mathbb{E}(M_\infty | \mathcal{F}_n) = \mathbb{E}(M_k | \mathcal{F}_n) + \mathbb{E}(M_\infty - M_k | \mathcal{F}_n) \geqslant M_n + \mathbb{E}(M_\infty - M_k | \mathcal{F}_n).$$

Since

$$\lim_{n\to\infty} \mathbb{E}\Big(\big|\mathbb{E}(M_\infty - M_k | \mathcal{F}_n)\big|\Big) \leqslant \lim_{n\to\infty} \mathbb{E}\Big(\mathbb{E}\big(|M_\infty - M_k| \, \big| \mathcal{F}_n\big)\Big) = \mathbb{E}(|M_\infty - M_k|) \overset{k\to\infty}{\longrightarrow} 0,$$

we have $\mathbb{E}(M_\infty - M_k | \mathcal{F}_n) \overset{k\to\infty}{\longrightarrow} 0$ a.s. Since $\mathbb{E}(M_k | \mathcal{F}_n) \geqslant M_n$ by the submartingale property, we have $\mathbb{E}(M_\infty | \mathcal{F}_n) - M_n \leqslant \mathbb{E}(M_\infty - M_k | \mathcal{F}_n)$, and the claim follows by taking $k \to \infty$. If (M_n) is a martingale, replace $\geqslant$ by $=$ at the appropriate places. □

The following theorem is the announced statement that all UI Martingales have the form of Example (4.62). In words, it says that every UI martingale is the 'discovery process' by finer and finer σ-algebras of a single 'limiting' RV X.

(4.68) Theorem

a) Let (M_n) be a UI-martingale. Then there exists a filtration $(\mathcal{F}_n)$ and a RV X such that

$$M_n = \mathbb{E}(X|\mathcal{F}_n) \text{ a.s.} \quad \text{and} \quad \lim_{n\to\infty} M_n = X \text{ in } L^1 \text{ and a.s..}$$

b) Let X be a RV and $(\mathcal{F}_n)$ a filtration. Define $\mathcal{F}_\infty := \lim_{n\to\infty} \mathcal{F}_n := \sigma(\mathcal{F}_n : n \in \mathbb{N})$, and $M_n := \mathbb{E}(X|\mathcal{F}_n)$. Then (M_n) is an UI-martingale, and

$$\lim_{n\to\infty} M_n = \mathbb{E}(X|\mathcal{F}_\infty) \text{ in } L^1 \text{ and a.s.}$$

Proof: a) Choose $X = M_\infty = \lim_{n\to\infty} M_n$ (a.s. limit), and $\mathcal{F}_n = \sigma(m_k : k \leqslant n)$. The statement then is just a reformulation of Theorem (4.67).

b) (M_n) is a martingale (example (4.9 c)), and it is UI by Theorem (4.64). By Theorem (4.67) then there exists a RV M_∞ such that $M_n \to M_\infty$ a.s. and in L^1. It remains to show that $M_\infty = \mathbb{E}(X|\mathcal{F}_\infty)$.

We first treat the case where $X \geqslant 0$. Then $Y := \mathbb{E}(X|\mathcal{F}_\infty) \geqslant 0$, and $M_\infty \geqslant 0$ a.s. Consider the finite measures μ_1, μ_2 on $(\Omega, \mathcal{F}_\infty)$ with

$$\forall A \in \mathcal{F}_\infty : \qquad \mu_1(A) = \mathbb{E}(M_\infty 1_A), \quad \mu_2(A) = \mathbb{E}(Y 1_A).$$

We claim that for each $m \in \mathbb{N}$ and each $A \in \mathcal{F}_m$, $\mu_1(A) = \mu_2(A) = \mathbb{E}(X 1_A)$. To see this, note that for $n > m$ we have

$$\mathbb{E}(X1_A) = \mathbb{E}(\mathbb{E}(X1_A|\mathcal{F}_n)) = \mathbb{E}(1_A \mathbb{E}(X|\mathcal{F}_n)) = \mathbb{E}(1_A M_n) \overset{n\to\infty}{\longrightarrow} \mathbb{E}(1_A M_\infty) = \mu_1(A),$$

and

$$\mathbb{E}(X1_A) = \mathbb{E}(\mathbb{E}(X1_A|\mathcal{F}_\infty)) = \mathbb{E}(1_A \mathbb{E}(X|\mathcal{F}_\infty)) = \mu_2(A).$$

This means that $\mu_1(A) = \mu_2(A)$ for all A from the π-system $\bigcup_{n=1}^\infty \mathcal{F}_n$. The system

$$\mathcal{L} := \sigma\big(\{A \in \bigcup_{n=1}^\infty \mathcal{F}_n : \mu_1(A) = \mu_2(A)\}\big)$$

is a σ-algebra, hence a λ-system. Since obviously $\bigcup_{n=1}^\infty \mathcal{F}_n \subset \mathcal{L}$, by the π-λ-Theorem (1.20) we have $\mathcal{F}_\infty = \sigma(\bigcup_{n=1}^\infty \mathcal{F}_n) \subset \mathcal{L}$. So, indeed $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{F}_\infty$. (This is, once more, the proof of the very useful statement: if two measures agree on a $\cap$-stable generator of a σ-algebra, they agree on the σ-algebra.)

Now by choosing $A = 1_{\{M_\infty > Y\}} \in \mathcal{F}_\infty$, we find that $\mathbb{E}((M_\infty - Y)1_{\{M_\infty > Y\}}) = \mu_1(A) - \mu_2(A) = 0$, so $M_\infty \leqslant Y$ a.s. Conversely, choosing $A = 1_{\{M_\infty < Y\}}$ gives $M_\infty < Y$ a.s. The claim follows. □

The final topic of this section is a sort of converse to the last theorem. Instead of increasing the information about some RV X, we now decrease it. The resulting RVs then contain less and less information.

(4.69) Backwards Martingale Convergence Theorem

Let $X \in L^1$, and let $(\mathcal{G}_{-n})_{n\in\mathbb{N}}$ be a filtration, meaning that $\mathcal{G}_{-n} \subset \mathcal{G}_{-m}$ when $n \geqslant m$. Put

$$M_{-n} := \mathbb{E}(X|\mathcal{G}_{-n}), \qquad \mathcal{G}_{-\infty} := \bigcap_{n\in\mathbb{N}} \mathcal{G}_{-n}.$$

Then for each $N \in \mathbb{N}$, the process $(M_k)_{-N \leqslant k \leqslant 1}$ is a martingale, $M_{-\infty} := \lim_{n\to\infty} M_{-n}$ exists a.s. and in L^1, and $M_{-\infty} = \mathbb{E}(X|\mathcal{G}_{-\infty})$.

Proof: $(M_k)_{-N \leqslant k \leqslant 1}$ is a martingale by the tower property. The number of upcrossings $U_N(M)$ of the martingale $(M_k)_{-N \leqslant k \leqslant 1}$ must be a.s. bounded by the same argument as the one given in Lemma (4.29). Therefore, again by the same argument as for martingale convergence, $\lim_{N\to\infty} M_{-N}$ exists a.s. Since $(M_{-k})_{k\in\mathbb{N}}$ is UI, the convergence also holds in L^1. Now in the same way as in (4.68) we show that $M_{-\infty} = \mathbb{E}(X|\mathcal{G}_{-\infty})$. The details are left as an exercise. □

Finally, we re-prove the Strong Law of Large Numbers using martingale techniques. We are restricted to iid RVs, but as a bonus, we even get convergence in L^1.

(4.70) SLLN with Martingale Proof

Let (X_n) be iid integrable RVs and set $\mu := \mathbb{E}(X_1)$. Then

$$\frac{1}{n} S_n := \frac{1}{n} \sum_{i=1}^{n} X_k \overset{n\to\infty}{\longrightarrow} \mu \qquad \text{a.s. and in } L^1.$$

Proof: Let $\mathcal{G}_{-n} = \sigma(S_m : m \geqslant n)$, and $\mathcal{G}_{-\infty} = \bigcap_{n\in\mathbb{N}} \mathcal{G}_n$. In Example (3.30), you were asked to show that

$$\mathbb{E}(X_1|\mathcal{G}_{-n}) = \frac{1}{n} S_n.$$

Theorem (4.69) now implies that

$$\lim_{n\to\infty} \frac{1}{n} S_n = \lim_{n\to\infty} \mathbb{E}(X_1|\mathcal{G}_{-n}) \quad \text{exists a.s. and in } L^1.$$

To see that the limit is a.s. constant, let

$$L(\omega) := \limsup_{n\to\infty} \frac{1}{n} S_n(\omega) = \limsup_{n\to\infty} \frac{1}{n}(X_{k+1} + \ldots + X_n),$$

where the last equality holds for each fixed $k \in \mathbb{N}$. This shows that L is $\mathcal{T}$-measurable, with $\mathcal{T}$ the terminal σ-algebra, i.e. $\mathcal{T} = \bigcap_{k\in\mathbb{N}} \mathcal{T}_k$, and $\mathcal{T}_k = \sigma(X_k, X_{k+1}, \ldots)$. BY Kolmogorovs 0-1-law, $\mathbb{P}(L \geqslant c) \in \{0, 1\}$ for all $c \in \mathbb{R}$, which means that L is a.s. constant. Then $L = \mathbb{E}(L) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}(S_n) = \mu$. □